Data-based comparison of frequency analysis methods: A general framework

Benjamin Renard, K. Kochanek, M. Lang, F. Garavaglia, E. Paquet, L. Neppel, K. Najib, Julie Carreau, P. Arnaud, Y. Aubert, et al.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

Data-based comparison of frequency analysis methods: a general framework

B. Renard[(1)], K. Kochanek[(1,7)], M. Lang[(1)], F. Garavaglia[(2)], E. Paquet[(2)], L. Neppel[(3)], K. Najib[(3)], J. Carreau[(3)], P. Arnaud[(4)], Y. Aubert[(4)], F. Borchi[(5)], J.-M. Soubeyroux[(5)], S. Jourdain[(5)], J.-M. Veysseire[(5)], E. Sauquet[(1)], T. Cipriani[(1)] and A. Auffray[(6)]

(1) Irstea, UR HHLY Hydrology-Hydraulics, Lyon, France

(2) EDF-DTG, Grenoble, France

(3) University Montpellier II, UMR HydroSciences, Montpellier, France

(4) Irstea, UR OHAX, Aix-en-Provence, France

(5) Meteo-France, Direction de la Climatologie, Toulouse, France

(6) Meteo-France, Direction Interrégionale Centre-Est, Lyon, France

(7) Institute of Geophysics, Polish Academy of Sciences, Warsaw, Poland

Submitted for publication in Water Resources Research

October 2012

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

Abstract

An abundance of methods have been developed over the years to perform the frequency analysis (FA) of extreme environmental variables. Although numerous comparisons between these methods have been implemented, no general comparison framework has been agreed upon so far. The objective of this paper is to build the foundation of a data-based comparison framework, which aims at complementing more standard comparison schemes based on Monte Carlo simulations or statistical testing. This framework is based on the following general principles: (i) emphasis is put on the predictive ability of competing FA implementations, rather than their sole descriptive ability measured by some goodness-of-fit criterion; (ii) predictive ability is quantified by means of reliability indices, describing the consistency between validation data (not used for calibration) and FA predictions; (iii) stability is also quantified, i.e. the ability of a FA implementation to yield similar estimates when calibration data change; (iv) the necessity to subject uncertainty estimates to the same scrutiny as point-estimates is recognized, and a practical approach based on the use of the predictive distribution is proposed for this purpose. This framework is then applied to a case study involving 364 gauging stations in France, where 10 FA implementations are compared. These implementations correspond to the local, regional and local-regional estimation of Gumbel and Generalized Extreme Value distributions. Results show that reliability and stability indices are able to reveal marked difference between FA implementations. Moreover, the case study also confirms that using the predictive distribution to indirectly scrutinize uncertainty estimates is a viable approach, with distinct FA implementations showing marked differences in the reliability of their uncertainty estimates. The proposed comparison framework therefore constitutes a valuable tool to compare the predictive reliability of competing FA implementations, along with the reliability of their uncertainty estimates.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

1. Introduction

Frequency analysis (FA) of extremes is one of the cornerstones of hazard quantification and risk assessment. Its basic objective is to estimate the distribution of some environmental variable X. Such a distribution can be used to estimate the exceedance probability of a given value of X, or alternatively, to estimate the p-quantile of X (where p denotes the non-exceedance probability). The estimation of quantiles is of great importance since they are used to design civil engineering structures (e.g. dams, reservoirs, bridges) or to map hazard-prone areas where restrictions may be enforced (e.g. building restrictions in flood zones).

FA has been the subject of extensive research, yielding an abundance of approaches that can roughly be classified as follows:

- At-Site FA is a standard statistical analysis: parameters of a pre-specified distribution are estimated based on at-site observations of the variable X.

- Climate/Weather-informed at-site FA uses additional meteorological [e.g., weather type, *Garavaglia et al.*, 2010] or climatic [e.g. Interdecadal Pacific Oscillation IPO, *Micevski et al.*, 2006a] information. This family of methods stems from the observation that the distribution of X depends on some climate or weather state variable.

- Historical and paleoflood analyses are based on documentary sources or proxy data from e.g. sediment deposits. Such information is used to extend the record period from the last decades to several centuries (historical data) or millennia (paleoflood data). Specific statistical frameworks have been developed to treat such additional information [e.g. *Stedinger and Cohn*, 1986; *O'Connel et al.*, 2002; *Parent and Bernier*, 2003; *Naulet et al.*, 2005; *Reis and Stedinger*, 2005; *Neppel et al.*, 2010; *Payrastre et al.*, 2011].

- Regional Frequency Analysis (RFA) jointly uses data from several sites to perform the inference, which may improve the precision of estimates [see e.g. *Durrans and Kirby*, 2004; *Yu et al.*, 2004; *Overeem et al.*, 2008; *Kyselý et al.*, 2011 for recent examples]. Moreover, RFA allows estimating quantiles and related uncertainties at an ungauged site.

- Model-based FA (sometimes referred to as “continuous simulation methods”) uses a simulation model reproducing the main characteristics of the environmental variable [*Arnaud and Lavabre*, 1999, for rainfall; *Boughton and Droop*, 2003, for floods]. Quantiles are then directly derived from long series generated from the model.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

Within each of these families, a large number of variants exist, differing in e.g. the assumed parent distribution (e.g. Generalized Extreme Value (GEV), Log-Pearson), the parameter estimation approach (e.g. maximum likelihood (ML), moment), the definition of homogenous regions or the choice of the simulation model. To avoid ambiguity, the following terminology is systematically used in this paper: a "FA family" refers to any of the previously described families, while a specific variant within a family is referred to as a "FA implementation". For instance, the local estimation of a GEV distribution with (i) the ML approach, and (ii) the moments approach will be considered as two distinct FA implementations, belonging to the same FA family.

In practice, users may feel lost facing so many FA implementations. Consequently, national guideline documents for flood FA help practitioners in realizing their analyses with best-practice methods. Such documents were released e.g. in the UK [*Reed et al.*, 1999], in the US [*Interagency Advisory Committee on Water Data*, 1982], in Switzerland [*Spreafico et al.*, 2003] or in Australia [*Institution of Engineers Australia*, 1987].

In addition to these end-user-oriented guideline documents, a large number of comparative studies between competing FA implementations have been reported in the research literature (a non-exhaustive review will be proposed in section 2). However, as noted by *Bobee et al.* [1993], the comparison framework varies from one study to another. *Bobee et al.* therefore advocated "*a systematic approach to comparing distributions used in flood frequency analysis*", which is still not agreed upon to our knowledge.

Moreover, in recent years there has been a growing emphasis on the importance of quantifying and communicating uncertainties in FA implementations [e.g., *Hall et al.*, 2004; *Naulet et al.*, 2005; *Renard et al.*, 2006a; *Renard et al.*, 2006b; *Kysely*, 2008; *Lee and Kim*, 2008; *Hine and Hall*, 2010; *Lima and Lall*, 2010; *Neppel et al.*, 2010]. However, while most FA implementations include an evaluation of uncertainties, the question of the reliability of estimated uncertainties has received less attention in FA [but see e.g. *Kysely*, 2008; *Garavaglia et al.*, 2011, for recent exceptions]. Other fields of environmental sciences (e.g. weather forecasting [*Dawid*, 1984; *Atger*, 1999; *Gneiting et al.*, 2007] or hydrological modeling [*Hall et al.*, 2007; *Laio and Tamea*, 2007; *Thyer et al.*, 2009; *Renard et al.*, 2010]) have recognized the need to scrutinize uncertainty estimates.

The general objective of this paper is to build the foundation of a methodological framework devoted to the data-based comparison of FA implementations. This framework aims to

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

complement (but not replace) other comparison frameworks based for instance on Monte-Carlo simulations or statistical testing. Importantly, the framework we are proposing is built in order to meet the following requirements:

[R1] It should enable the inclusion of any FA implementation, whatever its family (at-site, regional, model-based etc.).

[R2] It should enable the comparison of estimated uncertainties.

This paper is organized as follows. Section 2 proposes a short review of commonly used comparison frameworks, and emphasizes the differences between them in terms of underlying objectives, advantages and limitations. Section 3 then describes the data-based comparison framework. In particular, section 3.2 proposes several indices to quantify the performance of competing FA implementations, and section 3.3 introduces the predictive distribution as an indirect way to compare uncertainty estimates. A case study based on 364 gauging stations in France illustrates the application of the comparison framework (section 4). Limitations are discussed in section 5, before summarizing the main conclusions in section 6.

2. A short review of standard comparison frameworks

2.1. Simulation-based comparisons

Simulation-based approaches use Monte-Carlo-generated data. Knowing the true distribution, the performance of a FA implementation can be quantified by means of formal and objective statistical criteria such as bias, root mean squared error (RMSE), etc. This approach has been widely used for the comparison between various distributions and/or estimation approaches [e.g., *Hosking et al.*, 1985; *Kroll and Stedinger*, 1996; *Madsen et al.*, 1997a; *Madsen et al.*, 1997b; *Sankarasubramanian and Srinivasan*, 1999; *Durrans and Tomic*, 2001; *Ribatet et al.*, 2007; *He and Valeo*, 2009; *Meshgi and Khalili*, 2009], and for robustness studies [i.e., the performance of a method outside its conditions of application, see e.g., *Stedinger and Cohn*, 1986; *England et al.*, 2003b; *Markiewicz and Strupczewski*, 2009]. Moreover, an important advantage of simulation-based approaches is that they enable a formal evaluation of estimated uncertainties [see e.g. *Stedinger*, 1983b; *Stedinger and Tasker*, 1985; *Chowdhury and Stedinger*, 1991; *Cohn et al.*, 2001; *Kysely*, 2008; *Stedinger et al.*, 2008].

Simulation-based studies are hence useful, even necessary, to verify the internal consistency of a given FA implementation and to provide information about its main strengths and weaknesses. Indeed, a FA implementation performing poorly with synthetic data is unlikely to become highly capable with real data. Similarly, a FA implementation showing little

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

robustness with slight departures from its underlying assumptions should be considered with caution, since real data are unlikely to perfectly fulfill these assumptions.

However, good/better performance of a FA implementation with synthetic data is indicative only, but not conclusive, about its performance in practice. Indeed, determining whether the simulation setup is realistic enough to ensure that the good/better performances of a given FA implementation will also hold in real life is difficult. This is especially the case when FA implementations from distinct families are to be compared (requirement [R1]): for instance, deriving a simulation setup where local, regional and model-based FA implementations could be compared in a fair way is far from obvious.

2.2. Data-based comparisons

Data-based comparisons can complement simulation studies. Indeed, by using real data, they circumvent the difficulty of building realistic simulation setups. However, the main difficulty is that the truth is unknown, thus precluding the use of formal statistical criteria like bias or RMSE. Specific comparison schemes are therefore required. Data-based comparisons are mainly implemented using statistical tests and split-sample validation.

2.2.1. Statistical tests

A statistical test is used to evaluate whether observations can be considered as realizations from the assumed distribution family [e.g., *Chowdhury et al.*, 1991; *Laio*, 2004]. We stress that while this is an important question, choosing a distribution family is not the final objective of frequency analysis: indeed, even if the parent distribution family were known, the *estimated* distribution used for decision and design would still be affected by estimation errors, thus requiring further evaluation of its performance.

Statistical tests are hence useful to reject FA implementations that cannot be statistically reconciled with observations. Unfortunately, as noted by *Bobee et al.* [1993], such tests are not powerful with the typical sample size available for environmental data (usually hardly exceeding 50 elements). Consequently, it is often observed that several competing FA implementations cannot be rejected [*Laio*, 2004]. Again, this calls for alternative comparison approaches to attempt further distinguishing between such FA implementations.

Another difficulty is that statistical tests are simply not available for many FA implementations. This is problematic when FA implementations from distinct families are to be compared (requirement [R1]), since in general tests will be available for only a few of them. General-purpose testing procedures do exist [e.g. Cramer–von Mises or Anderson-

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

Darling tests, see *Stephens*, 1974], but in their standard form they compare observations with a fully specified distribution: this is not a realistic setting in frequency analysis where parameters are unknown and need to be inferred. Applying these tests in their standard form would systematically favor over-parameterized implementations. Consequently, specific corrections need to be implemented to account for estimation uncertainty, which is not an obvious task. As an illustration, *Laio* [2004] derived tests customized to extreme value distributions, but these tests are only applicable with particular estimators.

2.2.2. Split-sample evaluation

Another data-based comparison approach is based on the splitting of observations into a calibration (or estimation/training) set and a validation (or testing) set [*Gunasekara and Cunnane*, 1992]. This approach distinguishes between the descriptive and predictive abilities of FA implementations, which are two fundamentally distinct properties. The former refers to the ability of a FA implementation to *describe* past events used for parameter estimation (i.e., calibration events), whereas the latter refers to the ability to *predict* new events (i.e. validation events). While a FA implementation of poor descriptive ability has slim chance to become highly capable in predictive mode, the contrary is not true: a FA implementation that can provide a good description of calibration data may become inefficient in predictive mode.

Split-sample procedures have been mainly implemented for comparing regional FA implementations [e.g., *GREHYS*, 1996; *Grover et al.*, 2002; *Ouarda et al.*, 2006; *Neppel et al.*, 2007; *Szolgay et al.*, 2009]. The evaluation is usually achieved by comparing quantiles computed from validation sites (generally using an at-site estimate based on a long series) and quantiles given by the regional FA implementation (ignoring data at the validation site). Standard measures like bias or RMSE can then be used by considering locally-estimated quantiles as surrogate for the unknown true quantiles. While this may be acceptable for moderate quantiles when the record length at the validation site is large, it might become unrealistic for larger quantiles, which are affected by significant sampling errors. Further refinements of this general approach have been proposed [in particular, see the Bootstrap-based scheme implemented by *GREHYS*, 1996].

Split-sample comparisons have also been attempted and discussed for local FA implementations [see in particular *Beard*, 1974; *Interagency Advisory Committee on Water Data*, 1982; *Gunasekara and Cunnane*, 1992; *Garavaglia et al.*, 2011], but far less frequently

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

than for regional FA. This is because each series has to be decomposed into calibration and validation periods for local FA implementations, which requires using very long series.

Split-sample procedures are of interest because they compare FA implementations in the context they are designed for, where the objective is to predict upcoming events ("How should a dam be designed to ensure that it will withstand *upcoming* floods?"), as opposed to describe past event ("How should a dam be designed to ensure that it would have withstood *observed* floods?"). Moreover, they use FA implementations in operational-like conditions, where both model errors (i.e. misspecified distribution) and estimation errors coexist. We stress the difference between this objective and the objective behind statistical tests (identifying the parent distribution, or at least rejecting inappropriate ones).

Unfortunately, split-sample procedures are challenging to apply for two main reasons: (i) as in any data-based procedure, the truth is unknown; (ii) they require a large amount of data to be of any practical interest.

3. A data-based comparison framework

3.1. Notation and basic hypotheses

The data-based framework described in this section follows the path of split-sampling evaluation as described in previous section 2.2.2. Let X be the variable whose distribution is sought. It is assumed that a (large) dataset of observations from X is available, denoted by $\boldsymbol{x} = \left(x_k^{(i)} \right)_{i=1:N_{site},\, k=1:n^{(i)}}$. The superscript $^{(i)}$ denotes the site, the subscript $_k$ denotes the time step. Note that the number of observations at each site does not need to be identical. Using a similar notation, we denote by $\boldsymbol{c}$ the subset of $\boldsymbol{x}$ used for calibration, and $\boldsymbol{v}$ the complementary subset used for validation. In cases where no distinction is needed, we use the generic notation $\boldsymbol{d}$ to denote any one of $\boldsymbol{c}$ or $\boldsymbol{v}$.

The cumulative distribution function (cdf) of the unknown parent distribution of X at site i is denoted by $F^{(i)}$. A given FA implementation M makes an assumption on the distribution of X, yielding a cdf $F_M^{(i)}(y \mid \boldsymbol{\theta})$. In this notation, y is the value at which the cdf is evaluated and $\boldsymbol{\theta}$ represents a vector of unknown parameters. In most cases the distributional assumption is explicit, but it may also be implicit in the case of model-based implementations (e.g. for floods $F_M^{(i)}(y \mid \boldsymbol{\theta})$ would result from the rainfall-runoff transformation encapsulated in the hydrologic model). Parameter estimation is then performed by the FA implementation, yielding a particular parameter value $\hat{\boldsymbol{\theta}}$. The estimated distribution is then defined by

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

$\hat{F}_M^{(i)}(y) = F_M^{(i)}(y \mid \hat{\boldsymbol{\theta}})$. We stress the naming and notational distinction that will be consistently used throughout this paper between the *parent* distribution ($F^{(i)}$, the unknown distribution that generated observations), the *assumed* distribution ($F_M^{(i)}$, with unknown parameters) and the *estimated* distribution ($\hat{F}_M^{(i)}$ corresponding to the assumed distribution with one particular parameter value).

The performance indices defined in the next sections can be used for comparison under the following minimal hypotheses:

[H1] "Extremes" correspond to large values.

[H2] At-site data are temporally independent.

Assumption [H1] states that large return periods are associated with large values which is the case for most environmental variables (e.g. flood, wind, precipitation, etc.). If this assumption does not hold (e.g. low flow analysis with annual minimum values), all indices can be readily modified to account for extremes in the left tail of the distribution.

Assumption [H2] is more stringent: while the assumption of serial independence can be deemed acceptable for variables related to extreme localized events (e.g. storm winds, heavy rainfalls, floods, [*Pujol et al.*, 2007]), other variables sometimes exhibit significant serial dependence [e.g. *Hamed and Rao*, 1998; *Cohn and Lins*, 2005; *Koutsoyiannis*, 2010]. A detailed analysis of the effect of serial dependence on the comparison framework lies well beyond the scope of this paper. It is therefore assumed that data can be considered as serially independent, either because physical or empirical evidence suggests so or thanks to some data pre-processing (e.g. data sub-sampling).

3.2. Performance indices

3.2.1. Reliability and stability

The performance of competing FA implementations is judged according to two criteria: reliability and stability [*Garavaglia et al.*, 2011]. A reliable FA implementation yields an estimated distribution close to the (unknown) parent distribution, or in other words, it is able to assign correct exceedance probabilities. In practice, since the parent distribution is unknown, reliability has to be evaluated using observed data.

The stability of a FA implementation describes its ability to yield similar estimates when different data are used for calibration. In an industrial context, stable estimates are sought

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

when a whole group of structures is to be designed (e.g. power plants or dams fleet). Indeed, quantile estimates strongly varying with new observations would result in a frequent questioning of the design, which is problematic since a built structure cannot be continuously modified to track estimates' variability. Moreover, unstable estimates might cause the actual protection level to differ strongly from e.g. dam to dam, even if all dams are designed with the same target protection level.

It is stressed that both criteria do not play the same role in judging the performance of a FA implementation. In particular, stability cannot be used alone, because it does not give any information about the ability to predict observations (a FA implementation can be stable but totally unreliable). Consequently, reliability is assessed first in the comparison framework. When several FA implementations appear equally reliable, the additional insights provided by stability can be used to further discriminate between them.

3.2.2. Reliability: pval

This first reliability index aims to evaluate the overall agreement between the estimated distribution $\hat{F}_M^{(i)}$ and observations $d_k^{(i)}$ (either calibration or validation data can be used). For a given site *i* and time step *k*, it is defined as follows:

$$pval_k^{(i)} = \hat{F}_M^{(i)}(d_k^{(i)}) \quad (1)$$

Under the assumption that the estimation is reliable ($\hat{F}_M^{(i)} = F^{(i)}\ \forall i$), $pval_k^{(i)}$ are realizations from a uniform distribution on each site *i*: $pval_k^{(i)} \sim U[0;1]\ \forall i$ (see Appendix 1). Graphical diagnostics to assess the agreement between observed $\left(pval_k^{(i)}\right)_{k=1:n^{(i)}}$ and their theoretical distribution under the reliability hypothesis will be described in subsequent section 3.2.5.

The reliability assumption ($\hat{F}_M^{(i)} = F^{(i)}\ \forall i$) is worth commenting. It is quite clear that it will never be strictly met because of model and estimation errors. However, we use it as a working assumption, and we are looking in the data for evidence conflicting with it (which would materialize in the case of index *pval* by non-uniformly distributed values). This is the same rationale than that behind the use of a H0 hypothesis in statistical testing. However, we are only performing graphical diagnostics derived over an ensemble of sites here. While this allows making comparative statements on the relative reliability of FA implementations, it

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

does not provide a formal decision rule to reject the reliability assumption, as a statistical test would. The reason for this is discussed in subsequent section 5.3.

3.2.3. Reliability: N_T

The second reliability index is based on the number of exceedances of an estimated *T*-year quantile [e.g. *Interagency Advisory Committee on Water Data*, 1982, Appendix 14; *Gunasekara and Cunnane*, 1992; *Garavaglia et al.*, 2010]:

$$N_T^{(i)} = \sum_{k=1}^{n^{(i)}} 1_{[\hat{q}_T^{(i)};+\infty)}(d_k^{(i)}) \qquad \text{Where } 1_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

Under the reliability assumption ($\hat{q}_T^{(i)} = q_T^{(i)}$), $N_T^{(i)}$ is a realization from the binomial distribution: $N_T^{(i)} \sim Bin(n^{(i)}, 1/T)$ (see Appendix 1). As previously, dedicated graphical diagnostics will be described in subsequent section 3.2.5. Contrarily to index *pval* which quantifies the overall reliability, N_T focuses on reliability for prescribed *T*-year quantiles.

3.2.4. Reliability: FF

The index *FF*, used by e.g. *England et al.* [2003a] and *Garavaglia et al.* [2011], corresponds to the index *pval* computed on the maximum observed value of each site, $d_{\max}^{(i)}$:

$$FF^{(i)} = \hat{F}_M^{(i)}(d_{\max}^{(i)}) \tag{3}$$

Under the reliability assumption ($\hat{F}_M^{(i)} = F^{(i)}$), $FF^{(i)}$ is a realization from a Kumaraswamy distribution with parameters ($n^{(i)}$;1): $FF^{(i)} \sim K[n^{(i)};1]$, whose cdf can be written as follows (see Appendix 1):

$$F_K(t) = t^{n^{(i)}}, 0 \le t \le 1 \tag{4}$$

3.2.5. Graphical diagnostics based on reliability indices

Graphical diagnostics of reliability are based on the comparison between the reliability indices and their theoretical distribution under the reliability assumption. For a given site *i* with $n^{(i)}$ observations, let $z^{(i)}$ be any one of the indices defined in sections 3.2.2-3.2.4 (e.g. *FF*), and $H^{(i)}$ the cdf of its theoretical distribution under the reliability assumption (e.g. cdf of

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

a Kumaraswamy distribution $K(n^{(i)};1)$). A technical difficulty arises because $H^{(i)}$ depends on the number of observations $n^{(i)}$, which in general varies from site to site.

This issue can easily be overcome in the case of indices having a continuous cdf (namely, *pval* and *FF*) by using a probability-probability plot (pp-plot) representation: probability-transformed indices $H^{(i)}(z^{(i)})$ are plotted against empirical frequencies (see Figure 1a-c for illustrations). Under the reliability hypothesis, the probability-transformed values $H^{(i)}(z^{(i)})$ are indeed uniformly distributed between 0 and 1, irrespective of the number of observations $n^{(i)}$. Departures from the diagonal in the pp-plot have specific interpretations in terms of under/over-estimation or predictive failures (see Figure 1). Moreover, additional axis transformations might be valuable to focus on particular areas of the pp-plot. Typically, the pp-plot can be transformed into a Gumbel quantile-quantile plot (qq-plot) by applying the Gumbel quantile function to each axis (see Figure 1d-f for illustrations). This allows focusing on extreme values of indices. Note that any other continuous quantile function can be used, depending on the area of interest in the pp-plot.

The case of the discrete index N_T is more problematic, because the cdf of its theoretical binomial distribution is not continuous. Probability-transformed indices $H^{(i)}\left(N_T^{(i)}\right)$ are therefore not uniformly distributed. A possibility to overcome this difficulty is to randomize the values $H^{(i)}\left(N_T^{(i)}\right)$ in order to discard the cdf discontinuity induced by the discrete nature of the index N_T. This randomization is performed as follows. Let $b(-1) = 0$ and $b(j) = H^{(i)}(j) = \Pr(N \le j), j \ge 0$, where N is a random variable following a $Bin(n^{(i)}, 1/T)$ distribution. At a given site i, the value $N_T^{(i)}$ is transformed into probability space by randomly sampling a value $w^{(i)}$ from an uniform distribution between $b\left(N_T^{(i)}-1\right)$ and $b\left(N_T^{(i)}\right)$. This is to be compared with the non-randomized probability transformation, which corresponds to setting $w^{(i)} = b\left(N_T^{(i)}\right)$. This randomization ensures that the values $w^{(i)}$ are uniformly distributed between 0 and 1 under the reliability hypothesis (see Appendix 1). It is then possible to use the same pp-plot and qq-plot representations as discussed for continuous indices (see Figure 4d-e for illustrations).

3.2.6. Stability: $SPAN_T$

The stability of quantile estimates can be quantified by contrasting the values obtained with two different calibration datasets $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. The index $SPAN_T$ proposed by *Garavaglia et al.*

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

[2011] is used in this paper. It is a measure of the relative deviation between the two estimated *T*-year quantiles. Let $\hat{q}_T^{(i)}$ denotes the *T*-year quantile at site *i*, derived from the estimated distribution $\hat{F}_M^{(i)}$. For a given site *i*, $SPAN_T$ is defined as follows:

$$SPAN_T^{(i)} = \frac{\left|\hat{q}_T^{(i)}(\boldsymbol{c}_1) - \hat{q}_T^{(i)}(\boldsymbol{c}_2)\right|}{\frac{1}{2}(\hat{q}_T^{(i)}(\boldsymbol{c}_1) + \hat{q}_T^{(i)}(\boldsymbol{c}_2))} \quad (5)$$

The comparison between competing FA implementations can then be performed by comparing the distribution of $SPAN_T^{(i)}$ over all sites $i = 1{:}N_{site}$: the FA implementation whose $SPAN_T$ distribution remains the closest to zero is the most stable.

3.3. Comparing uncertainties

3.3.1. Motivation

One of the requirements for the comparison framework is to enable the comparison of estimated uncertainties. The term "*estimated* uncertainties" aims to emphasize the fact that uncertainty quantification depends on the assumptions underlying the FA implementation (e.g. distribution family, estimation approach, etc.). Consequently, there is a distinct possibility that such estimated uncertainties are unreliable if those assumptions are unrealistic [see e.g. the discussion by *Daly*, 2006 in the context of spatial interpolation methods].

Evaluating uncertainty estimates cannot be performed by counting the percentage of points inside a α% confidence interval in Figure 2, because the values on the *x*-axis are based on estimates of the exceedance probability (by means of a plotting position formula). As such, those values are affected by considerable uncertainties. The approach taken in this paper to circumvent this difficulty is to transform the uncertainty intervals shown in Figure 2 into a new distribution, named the predictive distribution.

This tool is well-known and widely used in Bayesian statistics [e.g. *Gelman et al.*, 1995] and hydrologic modeling [e.g. *Todini and Mantovan*, 2007; *Thyer et al.*, 2009; *Renard et al.*, 2010], and has also been proposed for FA applications [*Coles*, 2001, chapter 9; *Cox et al.*, 2002; *Meylan et al.*, 2008, chapter 7]. Moreover, the notion of "expected probability" discussed by e.g. *Stedinger* [1983a], *Rosbjerg and Madsen* [1998] or *Kuczera* [1999] is conceptually related to the notion of predictive distribution. The main advantage of this approach is that the methodology used to compare estimated distributions (red line in Figure 2) can be applied to predictive distributions, hence indirectly comparing estimated

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

uncertainties. This section defines the predictive distribution in both Bayesian and non-Bayesian contexts.

3.3.2. The Bayesian predictive distribution

Following the notation introduced in section 3.1, we use $f_M(y|\boldsymbol{\theta})$ and $\hat{f}_M(y)$ to denote the probability density function (pdf) of assumed and estimated distributions, respectively.

In Bayesian statistics, parameter inference is performed using the posterior distribution $p_M(\boldsymbol{\theta}|\boldsymbol{c})$, where $\boldsymbol{c}$ represents the calibration data. The predictive distribution of a future observable y given observed data $\boldsymbol{c}$ is defined by the following pdf [e.g. *Gelman et al.*, 1995]:

$$\hat{\pi}_M(y) = p_M(y|\boldsymbol{c}) = \int f_M(y|\boldsymbol{\theta})\, p_M(\boldsymbol{\theta}|\boldsymbol{c})\, d\boldsymbol{\theta} \qquad (6)$$

The predictive distribution $\hat{\pi}_M(y)$ hence corresponds to integrating the assumed distribution $f_M(y|\boldsymbol{\theta})$ over the posterior distribution of $\boldsymbol{\theta}$, $p_M(\boldsymbol{\theta}|\boldsymbol{c})$, which represents the uncertainty in estimating $\boldsymbol{\theta}$. By contrast, the estimated pdf $\hat{f}_M(y)$ corresponds to using the assumed distribution $f_M(y|\boldsymbol{\theta})$ for a fixed value $\hat{\boldsymbol{\theta}}$ of its parameters (most commonly the posterior mean, median or mode), hence ignoring estimation uncertainty. Figure 2 illustrates the difference between the predictive distribution $\hat{\pi}_M(y)$ and the estimated distribution $\hat{f}_M(y)$, and compares these distributions with the uncertainty bounds.

In practice, the integration in equation (6) cannot be performed analytically in general and has to be approximated numerically. Given that the posterior distribution $p_M(\boldsymbol{\theta}|\boldsymbol{c})$ is often explored using Markov Chain Monte Carlo (MCMC) samplers, such numerical approximation is usually implemented using a Monte Carlo scheme (see Appendix 2).

3.3.3. Non-Bayesian predictive distributions

The predictive distribution in equation (6) is not defined in a non-Bayesian context, because the posterior distribution $p_M(\boldsymbol{\theta}|\boldsymbol{c})$ does not exist in frequentist statistics, where $\boldsymbol{\theta}$ is considered as a non-random quantity. However, the estimator of $\boldsymbol{\theta}$, noted $\hat{\boldsymbol{\theta}}(X)$, is a random variable and its distribution is defined - it is the sampling distribution of the estimator. Note the distinction between the (random) estimator $\hat{\boldsymbol{\theta}}(X)$ and the (non-random) estimated value $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{c})$ corresponding to the value taken by the estimator on the calibration sample.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

The question of deriving a non-Bayesian version of the predictive distribution in equation (6) has attracted a lot of attention amongst statisticians. This has led to the development of innovative (sometimes controversial) inference paradigms, in particular pivotal inference and fiducial probabilities [*Fisher*, 1930; *Dawid and Stone*, 1982; *Seidenfeld*, 1992; *Dawid and Wang*, 1993; *Barnard*, 1995; *Wang*, 2000; *Lawless and Fredette*, 2005; *Hannig et al.*, 2006], predictive likelihoods [*Hinkley*, 1979; *Butler*, 1986; *Bjornstad*, 1990] and H-likelihoods [*Lee and Nelder*, 1996; *Meng*, 2009].

Harris [1989] proposed a pragmatic approach: the posterior distribution in equation (6) is simply replaced by the sampling distribution of $\hat{\boldsymbol{\theta}}(X)$. Let $s_M(\boldsymbol{\tau}|\boldsymbol{\theta})$ denote the pdf of this sampling distribution evaluated at $\boldsymbol{\tau}$. Note that in general, the sampling distribution depends on the unknown true parameter value $\boldsymbol{\theta}$. A non-Bayesian version of equation (6) is then:

$$\pi_M^*(y|\boldsymbol{\theta}) = \int f_M(y|\boldsymbol{\tau}) s_M(\boldsymbol{\tau}|\boldsymbol{\theta}) d\boldsymbol{\tau} \tag{7}$$

Compared with equation (6), there is an additional difficulty in equation (7) since the true value $\boldsymbol{\theta}$ is still unknown. *Harris*' proposal is to replace the unknown $\boldsymbol{\theta}$ by its estimated value $\hat{\boldsymbol{\theta}}$, yielding the following predictive distribution:

$$\hat{\pi}_M(y) = \pi_M^*(y|\hat{\boldsymbol{\theta}}) = \int f_M(y|\boldsymbol{\tau}) s_M(\boldsymbol{\tau}|\hat{\boldsymbol{\theta}}) d\boldsymbol{\tau} \tag{8}$$

Replacing the unknown true value $\boldsymbol{\theta}$ by its estimated value $\hat{\boldsymbol{\theta}}$ is a standard practice when estimating a sampling distribution. This is akin to the Fisher information matrix being replaced by the observed information matrix in ML estimation [e.g. *Coles*, 2001].

The predictive distribution in equation (8) was named the "parametric bootstrap predictive distribution" by *Harris* [1989], and has been further developed by other authors [e.g., *Basu and Harris*, 1994; *Vidoni*, 1995; *Fushiki et al.*, 2005; *Fushiki*, 2010]. A similar approach termed "bagging predictors" [e.g. *Breiman*, 1996] is used in the field of machine learning.

Similarly to the Bayesian predictive distribution, the integration in equation (8) in general is not performed analytically. Simple algorithms to derive non-Bayesian predictive distributions are described in Appendix 2. It is worth noting that deriving the predictive distribution only requires minimal effort beyond that made to quantify uncertainties.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

3.3.4. Indirectly comparing uncertainties via predictive distributions

The comparison of estimated uncertainties is then performed by replacing the cdf of the estimated distribution ($\hat{F}_M^{(i)}$) by the cdf of the predictive distribution $\hat{\Pi}_M(y)$ for all indices in section 3.2. The rationale behind this indirect approach is the following: if implementation *A* yields a more reliable predictive distribution than implementation *B* (according to the indices of section 3.2), it suggests that implementation A yields a more reasonable quantification of uncertainties in the sense that after transformation into a predictive distribution (eq. (6)-(8)), these uncertainties are in better agreement with validation data.

Throughout the remainder of this paper, we will simply use the naming "predictive distribution" with no further distinction between the Bayesian and the non-Bayesian versions. Indeed, while this distinction is necessary to introduce formal definitions, it is of little relevance in the context of the comparison framework discussed here.

4. Case study

The comparison framework described in previous sections is applied to a large runoff dataset. Ten FA implementations, belonging to three FA families, are compared. These implementations do not constitute an exhaustive representation of existing FA implementations, since the objective of this case study is not to draw definitive conclusions on the merits of existing FA implementations. Instead, it aims at illustrating the application of the performance indices described in section 3, and discussing the insights that can be gained from the application of a data-based comparison exercise.

4.1. Data and FA implementations

Daily runoff series from 364 stations in France are used (Figure 3), corresponding to catchment sizes ranging from 10 to 2,000 km^2. The time series cover at least 20 years, with more than 200 series spanning over 40 years. The quality of this dataset and its suitability for flood FA has been thoroughly evaluated in previous work [*Renard et al.*, 2008].

Annual maxima (AM) are extracted from the daily series. AM values are then treated with 10 FA implementations, belonging to three FA families, as summarized in Table 1:

1. Local estimation family: six implementations, corresponding to two distributional assumptions (Gumbel (GUM) and Generalized Extreme Value (GEV)) and three parameter estimation methods (Moments (MOM), Maximum Likelihood (ML) and Bayesian (BAY)), are used. The three estimation methods differ in their quantification of

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

uncertainty: (i) a non-parametric bootstrap approach is used for MOM; (ii) a standard Gaussian approximation for the sampling distribution of ML estimators is used; (iii) the posterior distribution of parameters represents the uncertainty in the Bayesian approach. For the latter approach, flat priors are used for location and scale parameters (i.e., $\pi(\theta) \propto 1$), while an Gaussian prior with mean 0 and standard deviation 0.2 is used for the shape parameter of the GEV distribution.

2. Regional estimation family: 2 implementations, corresponding to two distributional assumptions (GUM and GEV), are used. A standard index flood scheme [e.g. *Dalrymple*, 1960; *Robson and Reed*, 1999] is used: on the one hand, a regression between the index flood (taken here as the at-site mean) and catchment descriptors is built. On the other hand, a regional distribution is estimated by pooling standardized data (i.e. AM values divided by the index flood) from all sites together. Using the index flood regression together with the regional distribution enables estimating the distribution of AM at any site, including ungauged ones (see Appendix 3 for additional details).

3. Local-Regional estimation family: two implementations, corresponding to two distributional assumptions (GUM and GEV), are used. These implementations aim at using both the regional models above and the data observed at the target sites. The Bayesian approach proposed by *Ribatet et al.* [2006] is used: at each target site, the prediction by the regional model is used to define the prior distribution, while at-site data are used to build the likelihood function. The resulting posterior distribution therefore combines local and regional information (see Appendix 3 for details).

Note that the ten implementations analyzed in this case study correspond to fairly standard approaches, rather than state-of-the-art methods. Additional implementations could be considered to improve some aspects of the implementations described above. In particular, more advanced regionalization procedures could be investigated [e.g. *Madsen and Rosbjerg*, 1997; *Reis et al.*, 2005; *Micevski et al.*, 2006b; *Renard*, 2011]. However, we stress that the objective of this case study is not to provide the best possible estimation of flood quantiles in this particular area, but rather to illustrate the application of the comparison framework to standard FA implementations.

4.2. Reliability and stability decompositions

In order to assess the reliability of the FA implementations, the 364 sites are split into calibration and validation sets as follows:

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

- Sites with less than 40 years of data are used for calibration of the regional models (160 sites, red and pink dots in Figure 3a).
- For each site with more than 40 years of data (204 sites, black dots in Figure 3a), 20 years are randomly selected (independently from site to site) for calibration of the local models.
- For the latter sites, all remaining years (i.e. at least 20 years for each black dot in Figure 3a) are used for validation.

This decomposition allows comparing the reliability of all FA implementations based on exactly the same validation data.

Additional decompositions are required to assess stability. Since both local and regional implementations are considered, two types of decomposition are proposed:

- Stability with respect to local information (type I): for each site with more than 40 years of data (black dots in Figure 3a), two 20-year calibration sets are randomly selected. Purely regional implementations will be insensitive to this decomposition, since they do not use local information.
- Stability with respect to regional information (type II): sites with less than 40 years of data are split into two calibration sets (red and pink dots in Figure 3a). Purely local implementations will be insensitive to this decomposition.

4.3. Results

4.3.1. Illustration of the reliability diagnostics for one particular implementation

In order to illustrate the derivation of the graphical diagnostics of section 3.2.5, reliability is first evaluated for the sole implementation GEV_ML (the estimated distribution is used here). Figure 4a shows the pp-plot of *pval* for validation data, with each gray line corresponding to a validation site. Overall, the pp-curves are evenly distributed around the diagonal control line, and remain fairly close to it in most cases. However, the *pval* index only assesses the overall reliability, without particular focus on extremes. More stringent diagnostics are hence required to assess reliability at higher levels.

To this aim, Figure 4b shows the pp-plot of *FF*, for both calibration (blue) and validation (red) data. The S-shaped calibration curve indicates that the observed distribution of *FF* values is *less* variable than it would be if the parent distribution were used. This is an effect of errors in estimating GEV parameters, whose optimization tend to “over-fit” calibration data.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

At the opposite, the shape of the validation curve indicates that the distribution of *FF* values is *more* variable than it would be with the parent distribution. This implies that validation data are too often considered as "extreme" by the model, yielding high/low *FF* values with an unduly large frequency. In particular, a remarkable feature of this curve is its tendency to be stacked against the right border in the upper right corner: this corresponds to numerous *FF* values having *p*-values close to or equal to one, i.e. to observations that are considered as impossible by the model. This is a consequence of estimation errors for the shape parameter, yielding right-bounded GEV distributions whose bound is exceeded by validation data.

As suggested in section 3.2.5, an axis transformation can be used to focus on this area of the plot. Figure 4c therefore shows the same curves after transforming both axes into a Gumbel scale. Since this transformation is undefined for *FF* values equal to one, the corresponding points do not appear in the figure, but their percentage is reported. The large departure from the diagonal appearing in Figure 4c for the validation curve confirms the unduly high frequency of large *FF* values. In addition, 18% of validation data have a *FF* value equal to one – in other words, what is considered as impossible by the model actually occurs for 18% of the sites. This corresponds to severe prediction failures.

The second row of Figure 4 shows graphical diagnostics related to the N_{10} index. Figure 4d shows the N_{10} pp-plot after the randomization procedure described in section 3.2.5, while Figure 4e shows the qq-plot version of this diagnostic in Gumbel axes. These figures yield similar conclusions to the corresponding *FF* diagnostics.

The opposite behavior of calibration and validation curves is an illustration of the trade-off between descriptive and predictive capability: a too good fit to calibration data may come at the price of a reduced predictive reliability. In turn, this reemphasizes the necessity to assess predictive performances based on validation data. Consequently, all reliability diagnostics will focus on validation data in the remainder of this paper.

4.3.2. Comparison of estimation methods for local FA implementations

This section compares the three estimation methods (MOM, ML and BAY) used for local FA implementations. Figure 5 shows the reliability diagnostics for the estimated (first row, the posterior mode is used as parameter estimates) and the predictive (second row) distributions. For brevity, only the qq-plot representations in Gumbel space are reported, since it allows focusing on the most severe prediction failures.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

The *FF* diagnostic in Figure 5a indicates that the estimation method has little impact on reliability, compared to the choice of the distribution (GUM or GEV). Moreover, departures from the diagonal are smaller for the three GUM curves than for the three GEV curves. In addition, for MOM and ML estimation, validation data are considered as impossible by the GEV prediction for more than 15% of the sites. This percentage drops to 7% for BAY, which is a consequence of using an informative prior to constrain the shape parameter. The N_{10} and N_{100} diagnostics in Figure 5b-c yield similar insights. These results indicate that at-site estimation of a GEV distribution with 20 years of data may lead to substantial predictive failures, whatever the estimation method. This is a consequence of the well-documented difficulty in precisely identifying the shape parameter [e.g. *Coles*, 2001; *Garavaglia et al.*, 2011]. However, we stress that this does not imply that the GEV distribution should be rejected, but rather that local estimation with moderate sample size is not precise enough for this distribution to yield reliable predictions. In turn, this indicates that ignoring estimation uncertainty is not a viable option for locally-estimated GEV distributions.

The second row in Figure 5 shows the same diagnostics applied to the predictive distribution rather than the estimated distribution. The *FF* diagnostic in Figure 5d indicates that the estimation method (MOM, ML or BAY) has little impact for the Gumbel distribution: all three curves are similar and show marked departures below the diagonal. This suggests that even after accounting for uncertainty, a tendency to under-estimation remains with a Gumbel distribution. On the other hand, the estimation method has a stronger impact for the GEV distribution: the GEV_MOM curve shows the largest departure below the diagonal. Departure for the GEV_ML curve is similar although less pronounced. Lastly, the GEV_BAY curve appears much closer to the diagonal, suggesting that once uncertainties are accounted for, the predictions become fairly reliable. The N_{10} and N_{100} diagnostics in Figure 5e-f yield similar conclusions, although they reveal a more pronounced impact of the estimation method for the Gumbel distribution.

Overall, the results of this section suggest that while the estimation method does not strongly impact reliability based on the estimated distributions, the method used to quantify uncertainty exerts a stronger leverage according to the predictive distribution.

4.3.3. Comparison of local, regional and local-regional implementations

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

This section compares the three FA families (local, regional and local-regional) for both the Gumbel and the GEV distributions. For simplicity, only implementations GUM_BAY and GEV_BAY are used within the local family. Figure 6 shows the *pval* diagnostic for the estimated distribution, and for the three implementations involving the GEV distribution (similar plots are obtained with the Gumbel distribution, not shown). While local (GEV_BAY) and local-regional (GEV_LR) implementations yield similar diagnostics, the regional implementation (GEV_REG) shows marked departures from the diagonal for many sites. Such departures are more often below the diagonal (under-estimation) than above. Since the *pval* diagnostic does not focus on extremes, this indicates that the regional predictions may be markedly unreliable, even for small to moderate quantiles.

Figure 7 shows *FF*, N_{10} and N_{100} reliability diagnostics for the estimated distribution (first row) and the predictive distribution (second row). In Figure 7a (*FF*), the smallest departure from the diagonal corresponds to the GEV_LR implementation, while larger departures are observed for both regional implementations and the local GEV_BAY implementation. Figure 7b (N_{10}) highlights a clear distinction between regional implementations on the one hand, and local and local-regional implementations on the other hand. The former show a poor reliability even for predicting moderate 10-year quantiles, which confirms previous findings based on *pval*. On the other hand, local and local-regional implementations show similar predictive reliability for this index. However, the N_{100} index (Figure 7c) indicates that local-regional implementations become more reliable than local ones to predict larger quantiles. This suggests that while local approaches may be sufficient to estimate moderate quantiles, they become less reliable than local-regional approaches when extrapolated to higher quantiles, especially if a GEV distribution is used.

The second row of Figure 7 shows the same diagnostics applied to the predictive distribution. Figure 7d (*FF*) does not reveal marked differences between implementations, apart from GEV_BAY whose curve is closer to the diagonal as already observed in section 4.3.2. The N_{10} and N_{100} diagnostics (Figure 7e-f) yield more insights: in both cases, regional implementations show large departures from the diagonal, which suggests an unreliable quantification of uncertainty. On the other hand, local and local-regional implementations have similar curves for both indices, with smaller departures from the diagonal suggesting a more reliable quantification of uncertainty. The behavior of GEV_BAY for indices *FF* and N_{100} is noteworthy: while its predictions based on estimated distributions are unreliable (Figure 7a and c), it still yields fairly reliable predictions once uncertainty is accounted for

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

612 through the predictive distribution (Figure 7d and f). At the opposite, regional
613 implementations appear unreliable for both estimated and predictive distributions.

614 Lastly, stability is assessed by means of the $SPAN_{100}$ index (Figure 8). Figure 8a compares the
615 type-I stability of estimated (see section 4.2). Local implementations GEV_BAY and
616 GUM_BAY show the lowest stability. On the other hand, both local-regional
617 implementations GEV_LR and GUM_LR have similar stability, and importantly, are more
618 stable than any of the local implementations. Application of the $SPAN_{100}$ index to predictive
619 distributions yield identical insights (Figure 8c). Figure 8b compares the type-II stability of
620 the estimated distributions (see section 4.2). Both regional implementations GEV_REG and
621 GUM_ REG show a very low stability, while both local-regional implementations are far
622 more stable. Figure 8d shows a similar pattern for the predictive distribution.

623 Overall, the results of this section suggest that local-regional implementations generally
624 outperform both the purely local and regional implementations they are built upon. This
625 observation holds for both reliability (see e.g. Figure 7c) and stability (Figure 8).

626 # 5. Discussion

627 ## *5.1. Ability of reliability and stability indices to benchmark FA*
628 ## *implementations*

629 The case study shows that the indices defined in section 3.2 are able to reveal marked
630 difference between the reliability and stability of competing FA implementations. Moreover,
631 reliability indices appear quite complementary. Index *pval* is able to reveal reliability failures
632 at moderate levels (e.g. GEV_REG in Figure 6), but will not detect failures specific to
633 extreme levels. Index N_T allows focusing on specific quantiles, and varying the value of T
634 yields insights on the evolution of reliability at increasing levels (see e.g. the evolution of
635 GEV_BAY between N_{10} and N_{100} in Figure 7b-c). Lastly, index *FF* focuses on the most
636 extreme value observed at each site and is hence the most stringent reliability diagnostic. In
637 particular, it can reveal severe prediction failures, where observations as considered as
638 virtually impossible by the model (e.g. GEV_ML in Figure 4c). However, the set of indices
639 proposed in this paper is not exhaustive and could be completed in future work with
640 additional and possibly more powerful indices. As an illustration, *Garavaglia et al.* [2011]
641 assessed the stability of uncertainty estimates by quantifying the overlapping of confidence

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

intervals obtained using distinct calibration periods. Alternatively, indices based on the duration between exceedances of large quantiles could be derived as an alternative to N_T.

5.2. Feasibility of benchmarking uncertainty estimates

A major objective of this paper was to open uncertainty estimates to the same scrutiny as estimated distributions. This was achieved by transforming these uncertainties into a predictive distribution, which can be scrutinized in the same way as the estimated distribution. The results of the case study confirm that this is a viable approach, with distinct FA implementations showing marked differences in the reliability of their uncertainty estimates. Moreover, these results confirm two important points that are sometimes overlooked: (i) quantifying uncertainty is not sufficient, one also needs to assess whether this quantification is reliable [*Hall et al.*, 2007; *Thyer et al.*, 2009]; (ii) uncertainty estimates derived from a FA implementation whose assumptions are unrealistic are likely to be unreliable, and hence meaningless [*Daly*, 2006].

5.3. Limitations of the comparison framework

Despite showing its ability to compare FA implementations in terms of stability and reliability, the comparison framework remains based on a few hypotheses that are recalled and discussed in this section.

First, it is noted that the proposed framework only yields graphical comparisons between the stability and reliability of competing FA implementations. A natural extension would be to implement formal testing procedures, e.g. to test whether departures from the diagonal in Figure 5 are significant, or whether two FA implementations yield index distributions that are significantly different. Unfortunately, this is a challenging task because indices values are not independent from site to site, due to the spatial dependence between data. Consequently, the development of statistical tests would require a description of this spatial dependence. This was not attempted in this study because it would require making additional assumptions on the structure of spatial dependence beyond that made by the competing FA implementations.

Second, it is assumed that the data used for the comparison are temporally independent. Indeed, deriving the distribution of most performance indices (under the reliability hypothesis) requires making this assumption. It may be restrictive in some regions and/or for some hydrologic variables with significant inertia (e.g. low flows or mean annual runoff for groundwater-driven catchments). Future work could therefore evaluate the sensitivity of the comparison framework with temporally dependent data.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

Lastly, the general philosophy behind the comparison framework involves specific requirements, that are not limitations of the framework itself but might make its application difficult in some contexts. Indeed, applying the comparison framework requires an extended dataset of good-quality long series. The quality of the dataset needs to be thoroughly evaluated to avoid e.g. non-homogeneous data or heavily regulated catchments (see e.g. *Lang et al.* [2010] for examples of misleading results caused by the poor quality of a dataset). Moreover, the number of sites needs to be large enough since tools for assessing stability and reliability are based on the distribution of indices over an ensemble of sites. Lastly, long series are also required, since the evaluation of reliability remains limited by the series length: on the one hand, data left out for validation should be numerous to enable a truly challenging assessment of predictive ability; on the other hand, one needs to preserve enough data to calibrate the FA implementation.

A consequence of these requirements is that the comparison framework is geared toward large scale, national-wide comparisons rather than smaller-scale studies involving a couple of sites. In particular, the framework cannot compare predictive performance on one particular site. This is an acknowledged limitation, since a FA implementation having the best predictive performance on an ensemble of sites can still fail on one particular site.

5.4. Tailoring and developing comparison schemes

The case study of section 4 is performed at a rather large scale, which may restrict the ability to benchmark FA implementations. Indeed, it is likely that for daily runoff, the “best” FA implementation depends on various catchment properties like catchment size, elevation, climatic area, etc. Consequently, the comparison performed in this case study should be refined at the smaller scale of homogenous hydro-climatic regions. Moreover, the FA implementations compared in this case study are only a small sample of available FA implementations. More precisely, work is currently in progress to extend this comparison to additional distributions (e.g. log-Normal, Pearson family), estimation approaches (e.g. linear moments), approaches to uncertainty quantification (e.g. parametric bootstrap as advocated by *Kysely* [2008]), and alternative regionalization procedures.

Lastly, the decomposition into calibration and validation subsets could also be tailored to focus on more specific issues, for instance non-stationarity or low-frequency variability. As an illustration, the decomposition could be stratified according to the value of some climate

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

index (e.g. SOI, NAO) to evaluate the added value of implementations that use climate information.

5.5. Moving toward a systematic approach to comparing FA implementations

The data-based comparison approach presented in this paper might be a part of the systematic comparison approach advocated by *Bobee et al* [1993]. However it would not be reasonable to rely on a single comparison framework (let alone on a single comparison metric) to choose between competing FA implementations. We note that there have been controversies on which type of comparison framework should be used (see e.g. [*Wallis and Wood*, 1985; *Beard*, 1987; *Wallis and Wood*, 1987] for data-based vs. simulation studies). However, we claim that the different types of comparison frameworks should not be opposed, but rather be used together since they may actually yield complementary insights. Moreover, concordant results derived from distinct comparison frameworks constitute pieces of evidence that add up to build confidence in their generality [*Gunasekara and Cunnane*, 1992]. As an illustration, most results obtained in this paper are fully consistent with previous simulation-based studies, in particular the poor performance of the GEV distribution with small samples and no prior information [*Martins and Stedinger*, 2000] or the benefit of combining local and regional information [*Stedinger and Lu*, 1995]. The fact that similar findings are found in simulation-based and data-based contexts indicate that they can be extrapolated outside of the particular simulation setups used in the former comparisons.

Consequently, a comprehensive comparison of FA implementations might encompass the following steps:

1. Simulation studies in an “ideal” setup (no model misspecification) are useful to quantify the performance of FA implementations in formal statistical terms (e.g. bias, RMSE, reliable quantification of uncertainty, etc.). Moreover, “non-ideal” setups can be used to assess robustness. FA implementations that grossly fail this simulation step are probably not worth further investigation, but for other implementations, alternative comparison frameworks can provide a complementary point of view on their relative performance.

2. When available, statistical tests can be used to reject implementations that are in obvious disagreement with observations. An advantage of this comparison approach is that is can be implemented on a site-by-site basis. However, several implementations should be expected to pass the tests given their quite low power with typical sample sizes.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

3. The data-framework proposed in this paper evaluate the implementations' performance in terms of predictive ability, which closely corresponds to the context those implementations are designed for. However, implementing such a framework requires setting up extensive datasets, and conclusions can only be drawn for an ensemble of sites and can not be individually tailored for each site.

6. Conclusion

This paper proposes a general framework devoted to the data-based comparison of FA implementations. This framework is based on the following general principles:

- The performance of FA implementations is judged in terms of reliability and stability. The latter is evaluated in predictive mode, i.e. using data that are not used for calibration.
- The framework does not use any surrogate for the unknown true quantiles, but uses indices reflecting whether validation data are consistent with FA predictions.
- The necessity to scrutinize uncertainty estimates is recognized, and a practical solution based on the use of the predictive distribution is proposed.

The comparison framework is applied to a case study that uses 364 daily runoff series. The performances of ten FA implementations, belonging to three FA families, are compared. This case study demonstrates the ability of the comparison framework to benchmark FA implementations. Local-regional implementations were found to outperform both the purely local and regional implementations they are built upon, both in terms of reliability and stability. Marked differences were also found regarding the reliability of the predictive distribution, which confirms its relevance to indirectly compare uncertainty estimates.

Finally, although the comparison framework proposed in this paper proved its usefulness, it remains open to scrutiny and improvement. In particular, other stability and reliability indices could be defined, and comparison schemes could be tailored to specific regions or hydrologic variables. However, the general principles upon which the framework is built intend to be as general as possible. In particular, the importance of predictive reliability and the need to scrutinize uncertainty estimates are two points that hold to any FA implementation. Moreover, combining this data-based framework with alternative comparison schemes (e.g. based on Monte-Carlo simulations and statistical tests) is likely to yield complementary insights.

7. Acknowledgments

This work is funded by the French Research Agency (ANR) through the project EXTRAFLO (https://extraflo.cemagref.fr/). The HYDRO database (Ministry of environment) and EDF are gratefully acknowledged for providing the data. The helpful comments by Dan Rosbjerg, Jery Stedinger, three anonymous reviewers and the Associate Editor are gratefully acknowledged.

8. References

Arnaud, P., and J. Lavabre (1999), Using a stochastic model for generating hourly hyetographs to study extreme rainfalls, *Hydrological Sciences Journal*, *44*(3), 433-446.

Atger, F. (1999), The skill of ensemble prediction systems, *Monthly Weather Review*, *127*(9), 1941-1953.

Barnard, G. A. (1995), Pivotal Models and the Fiducial Argument, *Int. Stat. Rev.*, *63*(3), 309-323.

Basu, A., and I. R. Harris (1994), Robust Predictive-Distributions for Exponential-Families, *Biometrika*, *81*(4), 790-794.

Beard, L. R. (1974), *Flood Flow Frequency Techniques: A Report*, Center for Research in Water Resources.

Beard, L. R. (1987), Relative Accuracy of Log Pearson-Iii Procedures - Discussion, *Journal of Hydraulic Engineering-Asce*, *113*(9), 1205-1206.

Benichou, P., and O. Le Breton (1987), Prise en compte de la topographie pour la cartographie des champs pluviométriques statistiques, *La Météorologie*, *7*(19), 23-34.

Bjornstad, J. F. (1990), Predictive Likelihood: A Review, *Stat. Sci.*, *5*(1), 242-265.

Bobee, B., G. Cavadias, F. Ashkar, J. Bernier, and P. Rasmussen (1993), Towards a Systematic-Approach to Comparing Distributions Used in Flood Frequency-Analysis, *J. Hydrol.*, *142*(1-4), 121-136.

Boughton, W., and O. Droop (2003), Continuous simulation for design flood estimation - a review, *Environmental Modelling & Software*, *18*(4), 309-318.

Breiman, L. (1996), Bagging predictors, *Machine Learning*, *24*(2), 123-140.

Butler, R. W. (1986), Predictive Likelihood Inference with Applications, *J. R. Stat. Soc. Ser. B-Methodol.*, *48*(1), 1-38.

Chowdhury, J., and J. Stedinger (1991), Confidence Interval for Design Floods with Estimated Skew Coefficient, *Journal of Hydraulic Engineering*, *117*(7), 811-831.

Chowdhury, J. U., J. R. Stedinger, and L. H. Lu (1991), Goodness-of-Fit Tests for Regional Generalized Extreme Value Flood Distributions, *Water Resources Research*, *27*(7), 1765-1776.

Cipriani, T., T. Toilliez, and E. Sauquet (2012), Estimating 10 year return period peak flows and flood durations at ungauged locations in France, *La houille blanche; submitted*.

Cohn, T. A., and H. F. Lins (2005), Nature's style: Naturally trendy, *Geophys. Res. Lett.*, *32*(23).

Cohn, T. A., W. L. Lane, and J. R. Stedinger (2001), Confidence intervals for expected moments algorithm flood quantile estimates, *Water Resour. Res.*, *37*(6), 1695-1706.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, 210 pp., Springer-Verlag, London.

Cox, D. R., V. S. Isham, and P. J. Northrop (2002), Floods: some probabilistic and statistical approaches, *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, *360*(1796), 1389-1408.

Dalrymple, T. (1960), Flood frequency analyses, in *Water-supply paper 1543-A*, edited, US Geological Survey.

Daly, C. (2006), Guidelines for assessing the suitability of spatial climate data sets, *Int. J. Climatol.*, *26*(6), 707-721.

Dawid, A. P. (1984), Statistical Theory - the Prequential Approach, *J. R. Stat. Soc. Ser. A-Stat. Soc.*, *147*, 278-292.

Dawid, A. P., and M. Stone (1982), The Functional-Model Basis of Fiducial-Inference, *Annals of Statistics*, *10*(4), 1054-1067.

Dawid, A. P., and J. L. Wang (1993), Fiducial Prediction and Semi-Bayesian Inference, *Annals of Statistics*, *21*(3), 1119-1138.

Durrans, S. R., and S. Tomic (2001), Comparison of parametric tail estimators for low-flow frequency analysis, *J. Am. Water Resour. Assoc.*, *37*(5), 1203-1214.

Durrans, S. R., and J. T. Kirby (2004), Regionalization of extreme precipitation estimates for the Alabama rainfall atlas, *J. Hydrol.*, *295*(1-4), 101-107.

England, J. F., R. D. Jarrett, and J. D. Salas (2003a), Data-based comparisons of moments estimators using historical and paleoflood data, *J. Hydrol.*, *278*(1-4), 172-196.

England, J. F., J. D. Salas, and R. D. Jarrett (2003b), Comparisons of two moments-based estimators that utilize historical and paleoflood data for the log Pearson type III distribution, *Water Resources Research*, *39*(9).

Fisher, R. A. (1930), Inverse Probability, *Mathematical Proceedings of the Cambridge Philosophical Society*, *26*, 528-535.

Fushiki, T. (2010), Bayesian bootstrap prediction, *J. Stat. Plan. Infer.*, *140*(1), 65-74.

Fushiki, T., F. Komaki, and K. Aihara (2005), Nonparametric bootstrap prediction, *Bernoulli*, *11*(2), 293-307.

Garavaglia, F., J. Gailhard, E. Paquet, M. Lang, R. Garcon, and P. Bernardara (2010), Introducing a rainfall compound distribution model based on weather patterns sub-sampling, *Hydrol. Earth Syst. Sci.*, *14*(6), 951-964.

Garavaglia, F., M. Lang, E. Paquet, J. Gailhard, R. Garcon, and B. Renard (2011), Reliability and robustness of a rainfall compound distribution model based on weather pattern sub-sampling, *Hydrology and Earth System Sciences.*, *15*(2), 519-532.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian data analysis*, 526 pp., Chapman & Hall.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, *69*, 243-268.

GREHYS (1996), Inter-comparaison of regional flood frequency procedures for Canadian rivers., *J. Hydrol.*, *186*, 85-103.

Grover, P. L., D. H. Burn, and J. M. Cunderlik (2002), A comparison of index flood estimation procedures for ungauged catchments, *Can. J. Civ. Eng.*, *29*(5), 734-741.

Gunasekara, T. A. G., and C. Cunnane (1992), Split Sampling Technique for Selecting a Flood Frequency-Analysis Procedure, *J. Hydrol.*, *130*(1-4), 189-200.

Hall, J., E. O'Connell, and J. Ewen (2007), On not undermining the science: coherence, validation and expertise. Discussion of Invited Commentary by Keith Beven Hydrological Processes, 20, 3141-3146 (2006), *Hydrol. Process.*, *21*(7), 985-988.

Hall, M. J., H. F. P. van den Boogaard, R. C. Fernando, and A. E. Mynett (2004), The construction of confidence intervals for frequency analysis using resampling techniques, *Hydrol. Earth Syst. Sci.*, *8*(2), 235-246.

Hamed, K. H., and A. R. Rao (1998), A modified Mann-Kendall trend test for autocorrelated data, *J. Hydrol.*, *204*(1-4), 182-196.

Hannig, J., H. Iyer, and P. Patterson (2006), Fiducial generalized confidence intervals, *J. Am. Stat. Assoc.*, *101*(473), 254-269.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

Harris, I. R. (1989), Predictive Fit for Natural Exponential-Families, *Biometrika*, *76*(4), 675-684.

He, J. X., and C. Valeo (2009), Comparative Study of ANNs versus Parametric Methods in Rainfall Frequency Analysis, *J. Hydrol. Eng.*, *14*(2), 172-184.

Hine, D., and J. W. Hall (2010), Information gap analysis of flood model uncertainties and regional frequency analysis, *Water Resources Research*, *46*.

Hinkley, D. (1979), Predictive Likelihood, *Annals of Statistics*, *7*(4), 718-728.

Hosking, J. R. M., R. Wallis James, and F. Wood Eric (1985), An appraisal of the regional flood frequency procedure in the UK flood studies report, *Hydrological Sciences Journal*, *30*(1), 85-109.

Institution of Engineers Australia (1987), *Australian Rainfall and Runoff*, Engineers Australia.

Interagency Advisory Committee on Water Data (1982), *Guidelines for determining flood-flow frequency: Bulletin 17B of the Hydrology Subcommittee*, U.S. Geological Survey, Reston, Va.

Koutsoyiannis, D. (2010), HESS Opinions 'A random walk on water', *Hydrol. Earth Syst. Sci.*, *14*(3), 585-601.

Kroll, C. N., and J. R. Stedinger (1996), Estimation of moments and quantiles using censored data, *Water Resources Research*, *32*(4), 1005-1012.

Kuczera, G. (1999), Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference, *Water Resources Research*, *35*(5), 1551-1557.

Kysely, J. (2008), A Cautionary Note on the Use of Nonparametric Bootstrap for Estimating Uncertainties in Extreme-Value Models, *Journal of Applied Meteorology and Climatology*, *47*(12), 3236-3251.

Kyselý, J., L. Gaál, and J. Picek (2011), Comparison of regional and at-site approaches to modelling probabilities of heavy precipitation, *Int. J. Climatol.*, *31*(10), 1457-1472.

Laio, F. (2004), Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, *Water Resources Research*, *40*(9).

Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, *11*(4), 1267-1277.

Lang, M., K. Pobanz, B. Renard, E. Renouf, and E. Sauquet (2010), Extrapolation of rating curves by hydraulic modelling, with application to flood frequency analysis, *Hydrological sciences Journal.*, *55*(6), 883-898.

Lawless, J. F., and M. Fredette (2005), Frequentist prediction intervals and predictive distributions, *Biometrika*, *92*(3), 529-542.

Lee, K. S., and S. U. Kim (2008), Identification of uncertainty in low flow frequency analysis using Bayesian MCMC method, *Hydrol. Process.*, *22*(12), 1949-1964.

Lee, Y., and J. A. Nelder (1996), Hierarchical generalized linear models, *J. R. Stat. Soc. Ser. B-Methodol.*, *58*(4), 619-656.

Lima, C. H. R., and U. Lall (2010), Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow, *J. Hydrol.*, *383*(3-4), 307-318.

Madsen, H., and D. Rosbjerg (1997), Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling, *Water Resources Research*, *33*(4), 771-781.

Madsen, H., C. P. Pearson, and D. Rosbjerg (1997a), Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events .2. Regional modeling, *Water Resources Research*, *33*(4), 759-769.

909 Madsen, H., P. F. Rasmussen, and D. Rosbjerg (1997b), Comparison of annual maximum
910 series and partial duration series methods for modeling extreme hydrologic events .1. At-site
911 modeling, *Water Resources Research*, *33*(4), 747-757.
912 Mardhel, V., P. Frantar, J. Uhan, and A. Mio (2004), Index of development and persistence of
913 the river networks as a component of regional groundwater vulnerability assessment in
914 Slovenia., paper presented at Int. Conf. groundwater vulnerability assessment and mapping,
915 Ustron, Poland, 15-18 June 2004.
916 Markiewicz, I., and W. G. Strupczewski (2009), Dispersion measures for flood frequency
917 analysis, *Physics and Chemistry of the Earth*, *34*(10-12), 670-678.
918 Martins, E. S., and J. R. Stedinger (2000), Generalized maximum-likelihood generalized
919 extreme-value quantile estimators for hydrologic data, *Water Resources Research*, *36*(3), 737-
920 744.
921 Meng, X. L. (2009), Decoding the H-likelihood, *Stat. Sci.*, *24*(3), 280-293.
922 Meshgi, A., and D. Khalili (2009), Comprehensive evaluation of regional flood frequency
923 analysis by L- and LH-moments. II. Development of LH-moments parameters for the
924 generalized Pareto and generalized logistic distributions, *Stoch. Environ. Res. Risk Assess.*,
925 *23*(1), 137-152.
926 Meylan, P., A.-C. Favre, and A. Musy (2008), *Hydrologie fréquentielle: Une science*
927 *prédictive*, 173 pp., Presses polytechniques et universitaires romandes, Lausanne.
928 Micevski, T., S. W. Franks, and G. Kuczera (2006a), Multidecadal variability in coastal
929 eastern Australian flood data, *J. Hydrol.*, *327*(1-2), 219-225.
930 Micevski, T., G. Kuczera, and S. W. Franks (2006b), A Bayesian Hierarchical Regional Flood
931 Model, paper presented at 30th Hydrology and Water Resources Symposium, Engineers
932 Australia, Launceston, Tas, Australia, 4-7 Dec.
933 Naulet, R., M. Lang, T. B. M. J. Ouarda, D. Coeur, B. Bobee, A. Recking, and D. Moussay
934 (2005), Flood frequency analysis on the Ardeche river using French documentary sources
935 from the last two centuries, *J. Hydrol.*, *313*(1-2), 58-78.
936 Neppel, L., P. Arnaud, and J. Lavabre (2007), Extreme rainfall mapping: Comparison
937 between two approaches in the Mediterranean area, *C. R. Geosci.*, *339*(13), 820-830.
938 Neppel, L., B. Renard, M. Lang, P. A. Ayral, D. Coeur, E. Gaume, N. Jacob, O. Payrastre, K.
939 Pobanz, and F. Vinet (2010), Flood frequency analysis using historical data: accounting for
940 random and systematic errors, *Hydrol. Sci. J.-J. Sci. Hydrol.*, *55*(2), 192-208.
941 O'Connel, D. R. H., D. A. Ostenaa, D. R. Levish, and R. E. Klinger (2002), Bayesian flood
942 frequency analysis with paleohydrologic bound data, *Water Resources Research*, *38*(5).
943 Ouarda, T., J. M. Cunderlik, A. St-Hilaire, M. Barbet, P. Bruneau, and B. Bobee (2006),
944 Data-based comparison of seasonality-based regional flood frequency methods, *J. Hydrol.*,
945 *330*(1-2), 329-339.
946 Overeem, A., A. Buishand, and I. Holleman (2008), Rainfall depth-duration-frequency curves
947 and their uncertainties, *J. Hydrol.*, *348*(1-2), 124-134.
948 Parent, E., and J. Bernier (2003), Bayesian POT modeling for historical data, *J. Hydrol.*, *274*,
949 95-108.
950 Payrastre, O., E. Gaume, and H. Andrieu (2011), Usefulness of historical information for
951 flood frequency analyses: Developments based on a case study, *Water Resources Research*,
952 *47*.
953 Pujol, N., L. Neppel, and R. Sabatier (2007), Regional tests for trend detection in maximum
954 precipitation series in the French Mediterranean region, *Hydrol. Sci. J.-J. Sci. Hydrol.*, *52*(5),
955 956-973.
956 Reed, D. W., D. S. Faulkner, A. J. Robson, H. Houghton-Carr, and A. C. Bayliss (1999),
957 *Flood Estimation Handbook*, Institute of Hydrology, Wallingford.

Reis, D. S., and J. R. Stedinger (2005), Bayesian MCMC flood frequency analysis with historical information, *J. Hydrol.*, *313*(1-2), 97-116.
Reis, D. S., J. R. Stedinger, and E. S. Martins (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resources Research*, *41*(10).
Renard, B. (2011), A Bayesian Hierarchical Approach To Regional Frequency Analysis, *Water Resources Research*, *47*.
Renard, B., V. Garreta, and M. Lang (2006a), An application of Bayesian analysis and MCMC methods to the estimation of a regional trend in annual maxima, *Water Resources Research*, *42*(12).
Renard, B., M. Lang, and P. Bois (2006b), Statistical analysis of extreme events in a non-stationary context via a Bayesian framework., *Stoch. Environ. Res. Risk Assess.*, *21*, 97-112.
Renard, B., D. Kavetski, M. Thyer, G. Kuczera, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resources Research*, *46*.
Renard, B., et al. (2008), Regional methods for trend detection: Assessing field significance and regional consistency, *Water Resources Research*, *44*(8).
Ribatet, M., E. Sauquet, J. M. Gresillon, and T. B. M. J. Ouarda (2006), A regional Bayesian POT model for flood frequency analysis, *Stoch. Environ. Res. Risk Assess.*, *21*(4), 327-339.
Ribatet, M., E. Sauquet, J. M. Gresillon, and T. B. M. J. Ouarda (2007), Usefulness of the reversible jump Markov chain Monte Carlo model in regional flood frequency analysis, *Water Resources Research*, *43*(8).
Robson, A. J., and D. W. Reed (1999), *Flood Estimation Handbook. Volume 3: Statistical procedures for flood frequency estimation*, 338 pp., Wallingford.
Rosbjerg, D., and H. Madsen (1998), Design with uncertain design values, in *Hydrology in a Changing Environment, Vol III*, edited by H. Wheater and C. Kirby, pp. 155-163, John Wiley & Sons.
Sankarasubramanian, A., and K. Srinivasan (1999), Investigation and comparison of sampling properties of L-moments and conventional moments, *J. Hydrol.*, *218*(1-2), 13-34.
Seidenfeld, T. (1992), R.A. Fisher's Fiducial Argument and Bayes' Theorem, *Stat. Sci.*, *7*(3), 358-368.
Spreafico, M., R. Weingartner, M. Barben, A. Ryser, B. Hingray, A. Musy, and M. Niggli (2003), Evaluation des crues dans les bassins versants de Suisse*Rep.*, Département fédéral de l'environnement, des transports, de l'énergie et de la communication, Berne.
Stedinger, J., and L. Lu (1995), Appraisal of regional and index flood quantile estimators, *Stoch. Hydrol. Hydraul.*, *9*(1), 49-75.
Stedinger, J. R. (1983a), Design-Events with Specified Flood Risk, *Water Resources Research*, *19*(2), 511-522.
Stedinger, J. R. (1983b), Confidence-Intervals for Design-Events, *Journal of Hydraulic Engineering-Asce*, *109*(1), 13-27.
Stedinger, J. R., and G. D. Tasker (1985), Regional hydrologic analysis: 1. Ordinary, weighted and generalized least squares compared, *Water Resources Research*, *21*(9), 1421–1432 [Correction, Water Resour. Res.,1422(1425), 1844, 1986.].
Stedinger, J. R., and T. A. Cohn (1986), Flood Frequency-Analysis with Historical and Paleoflood Information, *Water Resources Research*, *22*(5), 785-793.
Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resources Research*, *44*.
Stephens, M. A. (1974), EDF Statistics for Goodness of Fit and Some Comparisons, *J. Am. Stat. Assoc.*, *69*(347), 730-737.

Szolgay, J., J. Parajka, S. Kohnova, and K. Hlavcova (2009), Comparison of mapping approaches of design annual maximum daily precipitation, *Atmos. Res.*, *92*(3), 289-307.
Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: a case study using bayesian total error analysis, *Water Resources Research*, *45*.
Todini, E., and P. Mantovan (2007), Comment on: 'On undermining the science?' by Keith Beven, *Hydrol. Process.*, *21*, 1633-1638.
Vidoni, P. (1995), A simple predictive density based on the p*-formula, *Biometrika*, *82*(4), 855-863.
Wallis, J. R., and E. F. Wood (1985), Relative Accuracy of Log Pearson-Iii Procedures, *Journal of Hydraulic Engineering-Asce*, *111*(7), 1043-1056.
Wallis, J. R., and E. F. Wood (1987), Relative Accuracy of Log Pearson-Iii Procedures - Closure, *Journal of Hydraulic Engineering-Asce*, *113*(9), 1210-1214.
Wang, Y. H. (2000), Fiducial intervals: What are they?, *American Statistician*, *54*(2), 105-111.
Wasson, J. G., A. Chandesris, H. Pella, and L. Blanc (2004), Les hydro-écorégions: une approche fonctionnelle de la typologie des rivières pour la directive cadre européenne sur l'eau, *Ingénieries*, *40*, 3-10.
Yu, P. S., T. C. Yang, and C. S. Lin (2004), Regional rainfall intensity formulas based on scaling property of rainfall, *J. Hydrol.*, *295*(1-4), 108-123.

9. Appendix 1: distribution of performance indices

9.1. Pval

Let $t \in [0;1]$. $\Pr(Pval_k^{(i)} \le t) = \Pr(\hat{F}_M^{(i)}(D_k^{(i)}) \le t)$

If the estimation is reliable ($\hat{F}_M^{(i)} = F^{(i)}$):

$$\begin{aligned} \Pr(Pval_k^{(i)} \le t) &= \Pr(F^{(i)}(D_k^{(i)}) \le t) \\ &= \Pr(D_k^{(i)} \le \{F^{(i)}\}^{-1}(t)) \\ &= F^{(i)}\left(\{F^{(i)}\}^{-1}(t)\right) = t \end{aligned}$$

This corresponds to the cdf of a uniform distribution on [0,1].

9.2. N_T

For a given time step *k*, the exceedance of a quantile $\hat{q}_T^{(i)}$ is a Bernoulli trial. If the estimation is reliable ($\hat{q}_T^{(i)} = q_T^{(i)}$), its success (meaning here the exceedance of the *T*-quantile) probability is 1/*T*. With the assumption of serial independence, the variable N_T therefore corresponds to the number of successes in $n^{(i)}$ independent Bernoulli experiments: its distribution is therefore Binomial, with parameters ($n^{(i)}$,1/*T*).

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

9.3. FF

Let $t \in [0;1]$. $\Pr(FF^{(i)} \le t) = \Pr(\hat{F}_M^{(i)}(D_{\max}^{(i)}) \le t)$

If the estimation is reliable ($\hat{F}_M^{(i)} = F^{(i)}$):

$$\begin{aligned}\Pr(FF^{(i)} \le t) &= \Pr(F^{(i)}(D_{\max}^{(i)}) \le t) \\ &= \Pr(D_{\max}^{(i)} \le \{F^{(i)}\}^{-1}(t)) \\ &= \Pr(D_k^{(i)} \le \{F^{(i)}\}^{-1}(t) \forall k = 1...n^{(i)}) \\ &= \left[F\left(\{F^{(i)}\}^{-1}(t)\right)\right]^{n^{(i)}} = t^{n^{(i)}}\end{aligned}$$

This corresponds to the cdf of the Kumaraswamy distribution with parameters ($n^{(i)}$,1). Note that the transition between lines 3 and 4 uses the serial independence hypothesis.

9.4. Randomized probability transformation for N_T

Let $W_T^{(i)}$ be a random variable whose distribution, conditional on $N_T^{(i)}$, is uniform between $b\left(N_T^{(i)} - 1\right)$ and $b\left(N_T^{(i)}\right)$ (see section 3.2.5). Recall that $b(j)$ is defined by $b(j) = \Pr(N \le j)$, with $N \sim Bin(n^{(i)}, 1/T)$. Let $t \in [0;1]$. The conditional cdf of $W_T^{(i)}$ is:

$$\Pr(W_T^{(i)} \le t \mid N_T^{(i)} = j) = \begin{cases} 0 & \text{if } t \le b(j-1) \\ \left[t - b(j-1)\right] / \left[b(j) - b(j-1)\right] & \text{if } b(j-1) \le t \le b(j) \\ 1 & \text{if } t \ge b(j) \end{cases}$$

The unconditional cdf of $W_T^{(i)}$ can then be derived by using the total probability law:

$$\Pr(W_T^{(i)} \le t) = \sum_{j=0}^{+\infty} \Pr(W_T^{(i)} \le t \mid N_T^{(i)} = j) \Pr(N_T^{(i)} = j)$$

Let k denote the integer verifying $b(k) \le t < b(k+1)$. The infinite sum above can then be decomposed as follows:

$$\begin{aligned}\Pr(W_T^{(i)} \le t) = &\sum_{j=0}^{k} \Pr(W_T^{(i)} \le t \mid N_T^{(i)} = j) \Pr(N_T^{(i)} = j) \\ &+ \Pr(W_T^{(i)} \le t \mid N_T^{(i)} = k+1) \Pr(N_T^{(i)} = k+1) \\ &+ \sum_{j=k+2}^{+\infty} \Pr(W_T^{(i)} \le t \mid N_T^{(i)} = j) \Pr(N_T^{(i)} = j)\end{aligned}$$

When $j \le k$, $b(j) \le b(k) \le t$, and $\Pr(W_T^{(i)} \le t \mid N_T^{(i)} = j) = 1$

When $j \ge k+2$, $t < b(k+1) \le b(j-1)$, and $\Pr(W_T^{(i)} \le t \mid N_T^{(i)} = j) = 0$

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

$$\text{Consequently, } \Pr(W_T^{(i)} \le t) = \sum_{j=0}^{k} 1 \times \Pr(N_T^{(i)} = j) + \frac{t - b(k)}{b(k+1) - b(k)} \Pr(N_T^{(i)} = k+1) + \sum_{j=k+2}^{+\infty} 0 \times \Pr(N_T^{(i)} = j)$$

1060
$$= \Pr(N_T^{(i)} \le k) + \frac{t - \Pr(N \le k)}{\Pr(N \le k+1) - \Pr(N \le k)} \Pr(N_T^{(i)} = k+1)$$

$$= \Pr(N_T^{(i)} \le k) + \frac{t - \Pr(N \le k)}{\Pr(N = k+1)} \Pr(N_T^{(i)} = k+1)$$

1061 Under the reliability hypothesis, $N_T^{(i)} \sim Bin(n^{(i)}, 1/T)$, which is the same distribution as that
1062 of N. Consequently, $\Pr(N \le k) = \Pr(N_T^{(i)} \le k)$ and $\Pr(N = k+1) = \Pr(N_T^{(i)} = k+1)$. The
1063 equation above therefore simplifies as follows:

1064
$$\Pr(W_T^{(i)} \le t) = \Pr(N_T^{(i)} \le k) + \frac{t - \Pr(N_T^{(i)} \le k)}{\Pr(N_T^{(i)} = k+1)} \Pr(N_T^{(i)} = k+1) = t$$

1065 This corresponds to the cdf of a uniform distribution between 0 and 1.

1066 10. Appendix 2: algorithms for predictive distributions

1067 *10.1. Bayesian predictive distributions*

1068 It is assumed that the Bayesian inference is performed using a Markov chain Monte Carlo
1069 (MCMC) sampler, yielding a sample $(\boldsymbol{\theta}^{(i)})_{i=1:N_{sim}}$ from the posterior distribution $p_M(\boldsymbol{\theta}|\boldsymbol{c})$.
1070 The pdf of the predictive distribution $\hat{\pi}_M(y)$ evaluated at y can then be approximated by:

$$\hat{\pi}_M(y) \approx \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} f_M(y|\boldsymbol{\theta}^{(i)}) \qquad (9)$$

1071 Note that it may be more practical to generate a large sample $(y^{(i)})_{i=1:N_{sim}}$ from the predictive
1072 distribution and use its empirical distribution as an approximation:

1073 Do i = 1:N_{sim}

1074 1. Sample $y^{(i)}$ from the distribution with pdf $f_M(y|\boldsymbol{\theta}^{(i)})$

1075 *10.2. Non-Bayesian predictive distributions*

1076 Let $\hat{s}_M(\boldsymbol{\tau})$ denote the pdf of the sampling distribution of the estimator $\hat{\boldsymbol{\theta}}(X)$. The non-
1077 Bayesian predictive distribution can be approximated using the same algorithms than in
1078 section 10.1, replacing the sample $(\boldsymbol{\theta}^{(i)})_{i=1:N_{sim}}$ from the posterior distribution by a sample
1079 $(\boldsymbol{\tau}^{(i)})_{i=1:N_{sim}}$ generated from the sampling distribution $\hat{s}_M(\boldsymbol{\tau})$.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

In practice, the algorithm used to generate the sample $(\boldsymbol{\tau}^{(i)})_{i=1:N_{sim}}$ depends on the way $\hat{s}_M(\boldsymbol{\tau})$ is derived. For instance, if bootstrap resampling of observations is used, a sample $(\boldsymbol{\tau}^{(i)})_{i=1:N_{sim}}$ is then available from the bootstrap replications of data. Alternatively, $\hat{s}_M(\boldsymbol{\tau})$ may be derived using a large-sample Gaussian approximation (as done in many estimation approaches) and whose generation poses no difficulty. In non-Gaussian approximation of $\hat{s}_M(\boldsymbol{\tau})$ and other complicated cases, specialized sampling algorithms (e.g. MCMC) may be required.

Finally, some FA implementations provide uncertainties expressed directly on quantiles rather than on parameters. In such a case, let $\hat{s}_{M,T}(q)$ denote the pdf of the sampling distribution of the estimated *T*-year quantile $\hat{Q}_T(X)$. A sample $(y^{(i)})_{i=1:N_{sim}}$ from the predictive distribution can be generated as follows:

Do $i = 1{:}N_{sim}$

1. Sample *u* from a uniform distribution on [0;1].
2. Compute $T = 1/(1\text{-}u)$
3. Sample $y^{(i)}$ from the sampling distribution of $\hat{Q}_T(X)$ with pdf $\hat{s}_{M,T}(q)$.

11. Appendix 3: Regional and local-regional FA implementations

11.1. Regional implementations based on an index flood model

Index flood regression: the index flood values at site *i*, v_i, are linked with catchment descriptors $w_i^{(1)},...,w_i^{(N_{cov})}$ using the following regression:

$$\log(v_i) = \beta_0 + \sum_{j=1}^{N_{cov}} \beta_j w_i^{(j)} + \varepsilon_i, \varepsilon_i \sim N(0,\sigma^2) \qquad (10)$$

Building on previous work by *Cipriani et al*. [2012], the following catchment descriptors are used: (i) catchment area; (ii) mean catchment elevation; (iii) mean of 10-year daily rainfall within the catchment, as estimated by *Benichou and Le Breton* [1987]; (iv) mean IDPR index. The latter index (Index of Development and Persistence of the River networks) was proposed by *Mardhel et al.* [2004] as an indicator of infiltration capacity. Moreover, region-specific regressions are estimated, with regions shown in Figure 3b and based on the Hydro-ecoregions defined by *Wasson et al.*[2004].

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

Estimation of regression parameters $\beta_0,...,\beta_{N_{cov}}$ and residual standard deviation σ is performed using a Bayesian approach (with flat priors on $\left(\beta_0,...,\beta_{N_{cov}},\log(\sigma)\right)$).

Regional distribution estimation: Depending on the FA implementation, regional Gumbel or GEV distributions are estimated in each region, based on all standardized data of the region pooled together. A Bayesian approach (with flat priors) is used.

Prediction at target site: At a target site *k*, the estimated distribution is a Gumbel or a GEV distribution (depending on the FA implementation) with parameters:

$$\begin{gathered}\text{location: } \mu_k = \hat{v}_k \times \hat{\mu}_{reg} \text{, with } \hat{v}_k = \exp\left[\hat{\beta}_0 + \sum_{j=1}^{N_{cov}} \hat{\beta}_j w_k^{(j)}\right] \\ \text{scale: } \lambda_k = \hat{v}_k \times \hat{\lambda}_{reg} \\ \text{shape (GEV distribution only): } \xi_k = \hat{\xi}_{reg}\end{gathered} \tag{11}$$

The predictive distribution is derived by propagating forward the MCMC samples of $\left(\beta_0,...,\beta_{N_{cov}},\sigma,\mu_{reg},\lambda_{reg},\xi_{reg}\right)$, as outlined in section 10.1. The MCMC algorithm used in this paper is described by *Renard et al.* [2006a].

11.2. Local-regional implementations

Propagating forward the MCMC samples of $\left(\beta_0,...,\beta_{N_{cov}},\sigma,\mu_{reg},\lambda_{reg},\xi_{reg}\right)$ into equation (11) yields a large number of replicates for the Gumbel (or GEV) parameters at the target site. These replicates can be used to specify a prior for the local-regional implementation. To this aim, a Gaussian distribution is estimated based on the replicates, and is used as the prior distribution for the local-regional implementation. The rest of the analysis then proceeds as in standard local implementations.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

1123 **List of captions**

1124 Table 1. Summary of the FA implementations studied in this paper.

1125 Figure 1. Typical shapes for pp-plots (a-c) and qq-plots in Gumbel space (d-f).

1126 Figure 2. Illustration of the difference between the estimated distribution (with pdf $\hat{f}_M(y)$)
1127 and the predictive distribution (with pdf $\hat{\pi}_M(y)$). This illustrative figure results from the
1128 Bayesian estimation of a GEV distribution using 25 observations. Uncertainty intervals are
1129 quantile posterior intervals.

1130 Figure 3. Location of the study sites. (a) Decomposition into "regional sites" used to estimate
1131 the regional models and "local sites" used for local estimation and validation; (b) Regions
1132 derived from the Hydro-ecoregions of *Wasson et al.* [2004].

1133 Figure 4. Reliability diagnostics applied to the implementation GEV-ML (local estimation of
1134 a GEV distribution with maximum likelihood). (a) pval pp-plot. Each gray line refers to a
1135 validation site. (b) FF pp-plot. Red = validation data, blue = calibration data. (c) FF qq-plot in
1136 Gumbel space. The percentages of "impossible observations" (i.e. observations incompatible
1137 with the estimated GEV, yielding FF=1) are provided. (d) Randomized pp-plot of N10
1138 computed on all available observations; (e) Randomized qq-plot of N10 in Gumbel space.

1139 Figure 5. Reliability diagnostics for the six local FA implementations. First row = estimated
1140 distribution, second row = predictive distribution. (a) and (d): FF qq-plot in Gumbel space; (b)
1141 and (e): N10 qq-plot in Gumbel space; (c) and (f): N100 qq-plot in Gumbel space.

1142 Figure 6. pval pp-plot for local, local-regional and regional estimation of the GEV distribution
1143 (estimated distribution).

1144 Figure 7. Reliability diagnostics for six FA implementations (local, regional and local-
1145 regional, with Gumbel and GEV distributions). First row = estimated distribution, second row
1146 = predictive distribution. (a) and (d): FF qq-plot in Gumbel space; (b) and (e): N10 qq-plot in
1147 Gumbel space; (c) and (f): N100 qq-plot in Gumbel space.

1148 Figure 8. Stability diagnostic for six FA implementations (local, regional and mixed local-
1149 regional, with Gumbel and GEV distributions). Left = type I decomposition, right = type II
1150 decomposition. (a) – (b) = estimated distribution, (c) – (d) = predictive distribution.

1151

Notation list

$\boldsymbol{x} = \left(x_k^{(i)}\right)_{i=1:N_{site},k=1:n^{(i)}}$ observations

$\boldsymbol{c}$ subset of $\boldsymbol{x}$ used for calibration

$\boldsymbol{v}$ subset of $\boldsymbol{x}$ used for validation

$\boldsymbol{d}$ denotes either one of $\boldsymbol{c}$ or $\boldsymbol{v}$

$F^{(i)}(y)$ cdf of the parent distribution (evaluated at some value y)

$F_M^{(i)}(y|\boldsymbol{\theta})$ cdf of the assumed distribution in implementation M, with unknown parameters $\boldsymbol{\theta}$

$f_M(y|\boldsymbol{\theta})$ pdf of the assumed distribution in implementation M, with unknown parameters $\boldsymbol{\theta}$

$\hat{F}_M^{(i)}(y)$ cdf of the estimated distribution in implementation M

$\hat{f}_M(y)$ pdf of the estimated distribution in implementation M

$\hat{IT}_M(y)$ cdf of the predictive distribution in implementation M

$\hat{\pi}_M(y)$ pdf of the predictive distribution in implementation M

$\hat{\boldsymbol{\theta}}(X)$ estimator of unknown parameters $\boldsymbol{\theta}$

$\hat{\boldsymbol{\theta}}$ estimated value of $\boldsymbol{\theta}$

$s_M(\boldsymbol{\tau}|\boldsymbol{\theta})$ pdf of the sampling distribution of $\hat{\boldsymbol{\theta}}(X)$ in implementation M (evaluated at some value $\boldsymbol{\tau}$)

$\hat{s}_M(\boldsymbol{\tau})$ pdf of the estimated sampling distribution of $\hat{\boldsymbol{\theta}}(X)$ in implementation M

$p_M(\boldsymbol{\theta}|\boldsymbol{c})$ posterior distribution of $\boldsymbol{\theta}$ given observations $\boldsymbol{c}$ in implementation M

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

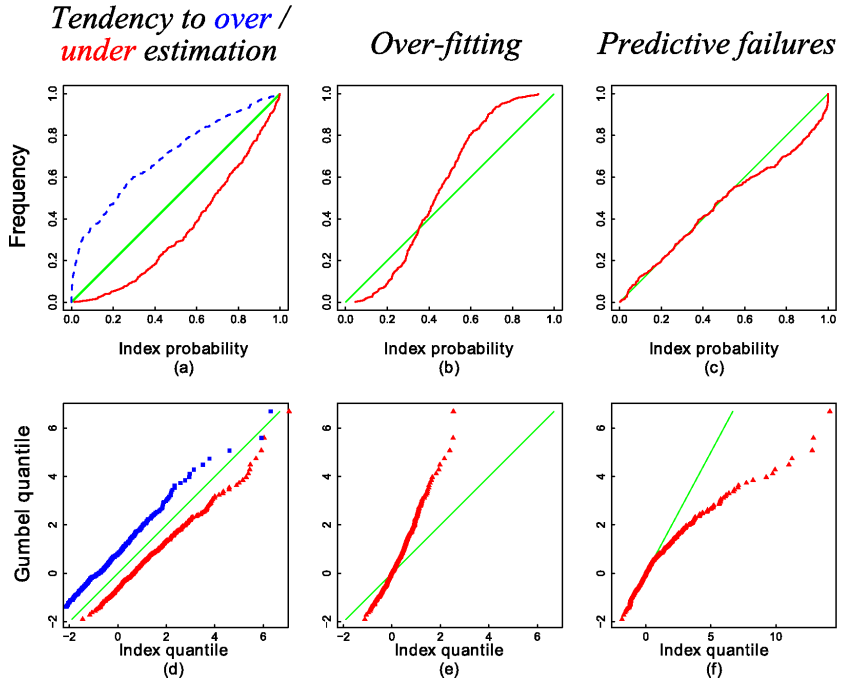
1169

1170 **Table 1. Summary of the FA implementations studied in this paper.**

Distribution	Estimation method	Notation	Uncertainty Quantification
FA Family: local estimation			
GEV	Moments	GEV_MOM	Bootstrap
GEV	Maximum Likelihood	GEV_ML	Gaussian approximation[1]
GEV	Bayesian	GEV_BAY	Bayesian
Gumbel	Moments	GUM_MOM	Bootstrap
Gumbel	Maximum Likelihood	GUM_ML	Gaussian approximation[1]
Gumbel	Bayesian	GUM_BAY	Bayesian
FA Family: regional estimation			
GEV	Bayesian	GEV_REG	Bayesian
Gumbel	Bayesian	GUM_REG	Bayesian
FA Family: local-regional estimation			
GEV	Bayesian	GEV_LR	Bayesian
Gumbel	Bayesian	GUM_LR	Bayesian

1171 [1]*Asymptotic normality of ML estimator, with covariance matrix equal to the Fisher*
1172 *information matrix.*

1173

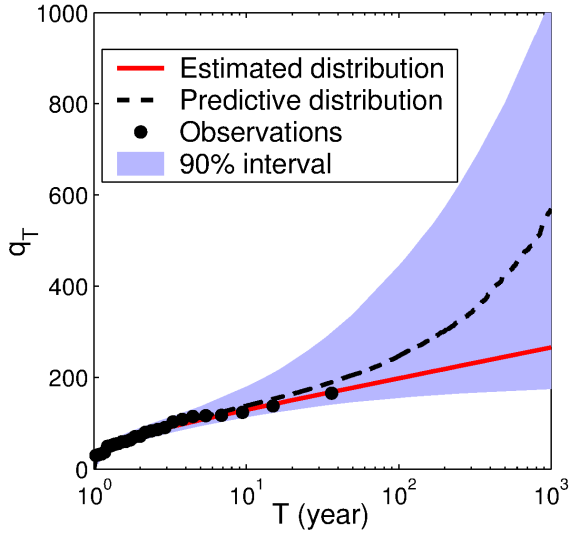


1174

1175 **Figure 1. Typical shapes for pp-plots (a-c) and qq-plots in Gumbel space (d-f).**

1176

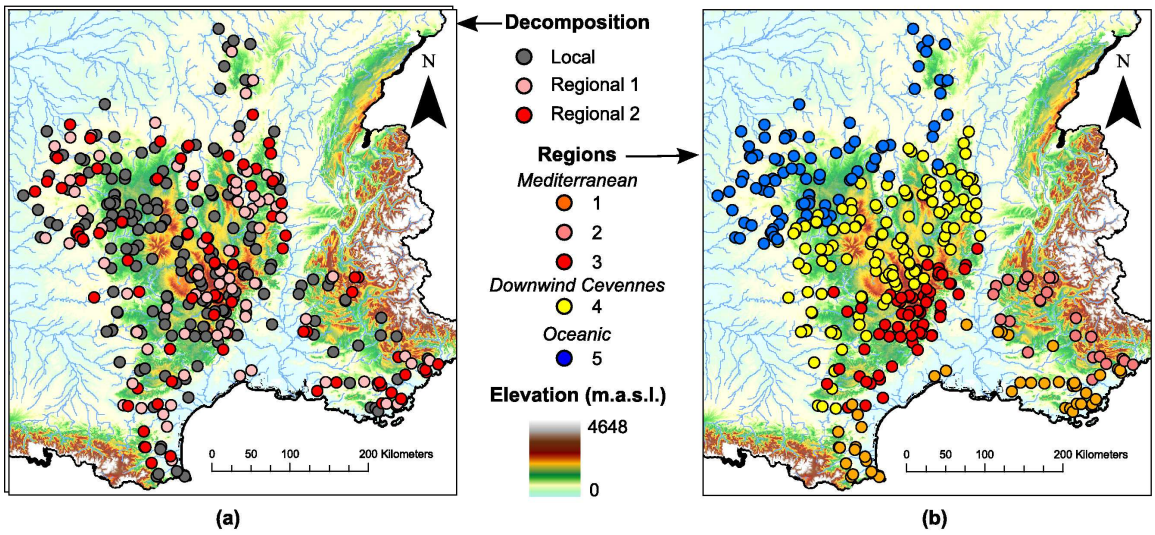
Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087



1177

1178 **Figure 2. Illustration of the difference between the estimated distribution (with pdf $\hat{f}_M(y)$) and the**
1179 **predictive distribution (with pdf $\hat{\pi}_M(y)$). This illustrative figure results from the Bayesian estimation of**
1180 **a GEV distribution using 25 observations. Uncertainty intervals are quantile posterior intervals.**

1181



1182

1183 **Figure 3. Location of the study sites. (a) Decomposition into "regional sites" used to estimate the regional**
1184 **models and "local sites" used for local estimation and validation; (b) Regions derived from the Hydro-**
1185 **ecoregions of *Wasson et al.* [2004].**

1186

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

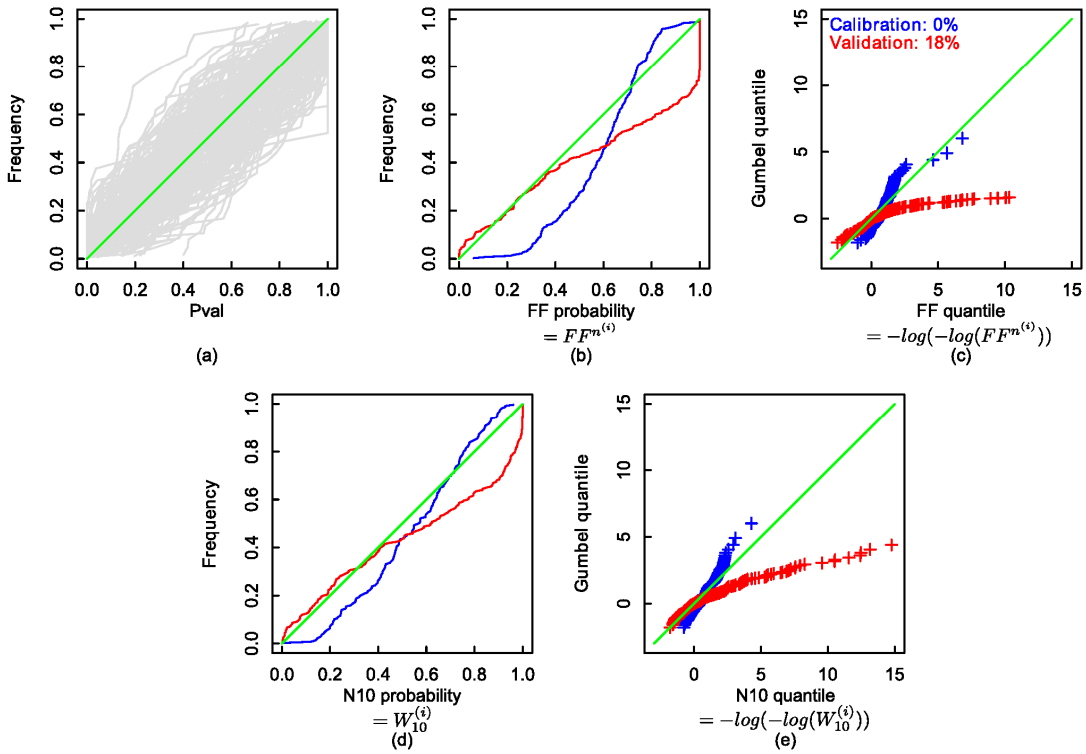


Figure 4. Reliability diagnostics applied to the implementation GEV-ML (local estimation of a GEV distribution with maximum likelihood). (a) *pval* pp-plot. Each gray line refers to a validation site. (b) *FF* pp-plot. Red = validation data, blue = calibration data. (c) *FF* qq-plot in Gumbel space. The percentages of "impossible observations" (i.e. observations incompatible with the estimated GEV, yielding *FF*=1) are provided. (d) Randomized pp-plot of N_{10} computed on all available observations; (e) Randomized qq-plot of N_{10} in Gumbel space.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

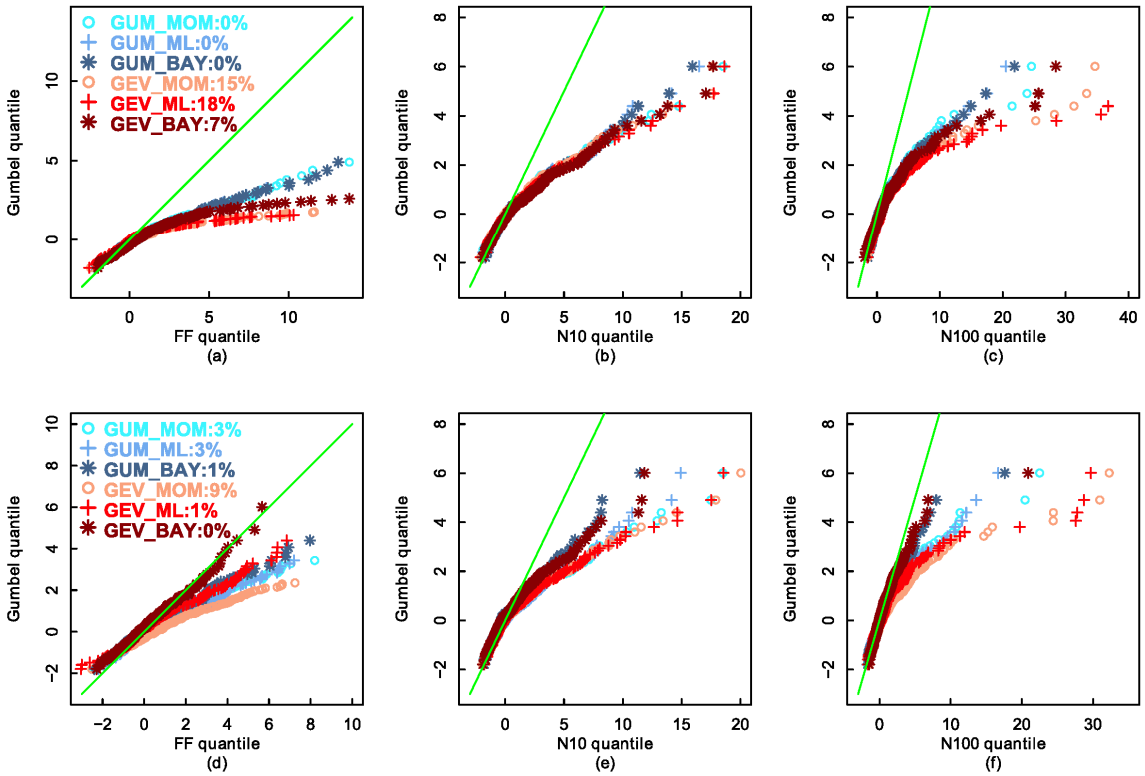


Figure 5. Reliability diagnostics for the six local FA implementations. First row = estimated distribution, second row = predictive distribution. (a) and (d): *FF* qq-plot in Gumbel space; (b) and (e): N_{10} qq-plot in Gumbel space; (c) and (f): N_{100} qq-plot in Gumbel space.

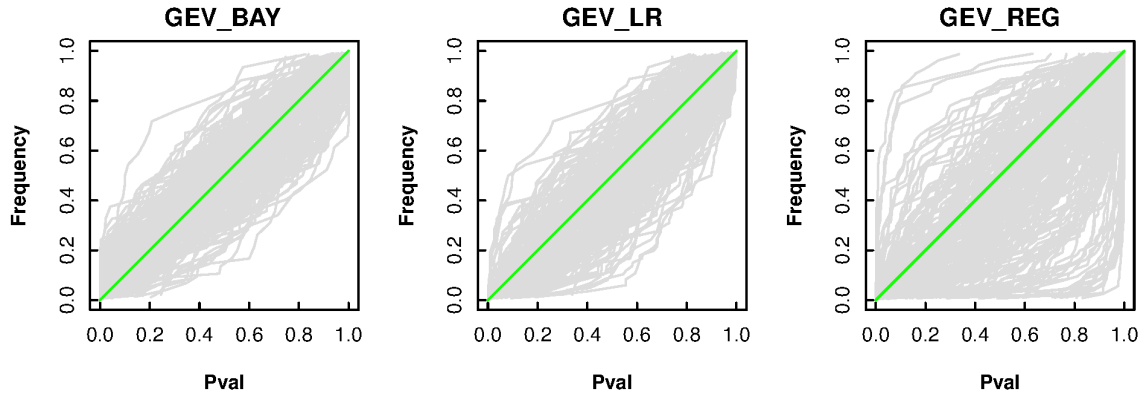


Figure 6. *pval* pp-plot for local, local-regional and regional estimation of the GEV distribution (estimated distribution).

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

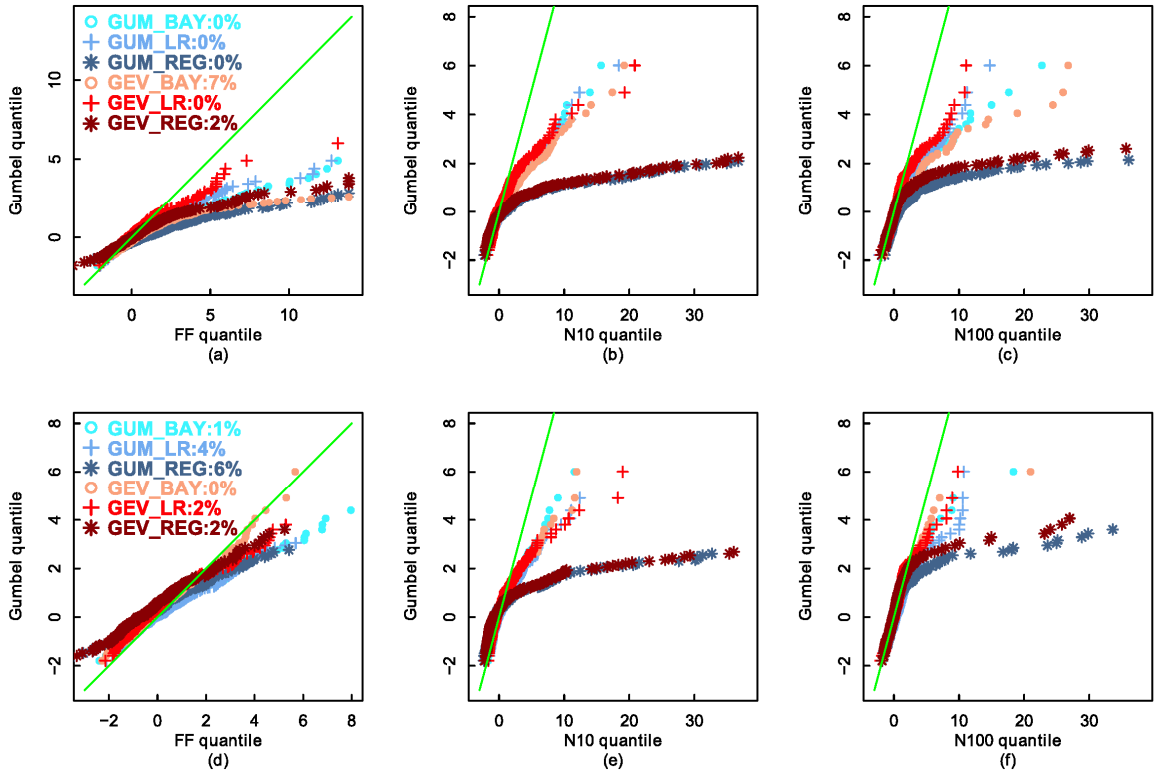


Figure 7. Reliability diagnostics for six FA implementations (local, regional and local-regional, with Gumbel and GEV distributions). First row = estimated distribution, second row = predictive distribution. (a) and (d): *FF* qq-plot in Gumbel space; (b) and (e): N_{10} qq-plot in Gumbel space; (c) and (f): N_{100} qq-plot in Gumbel space.

Author-produced version of the article published in Water Resources Research, vol. 49, 2013
The original publication is available at http://onlinelibrary.wiley.com/ DOI: 10.1002/wrcr.20087

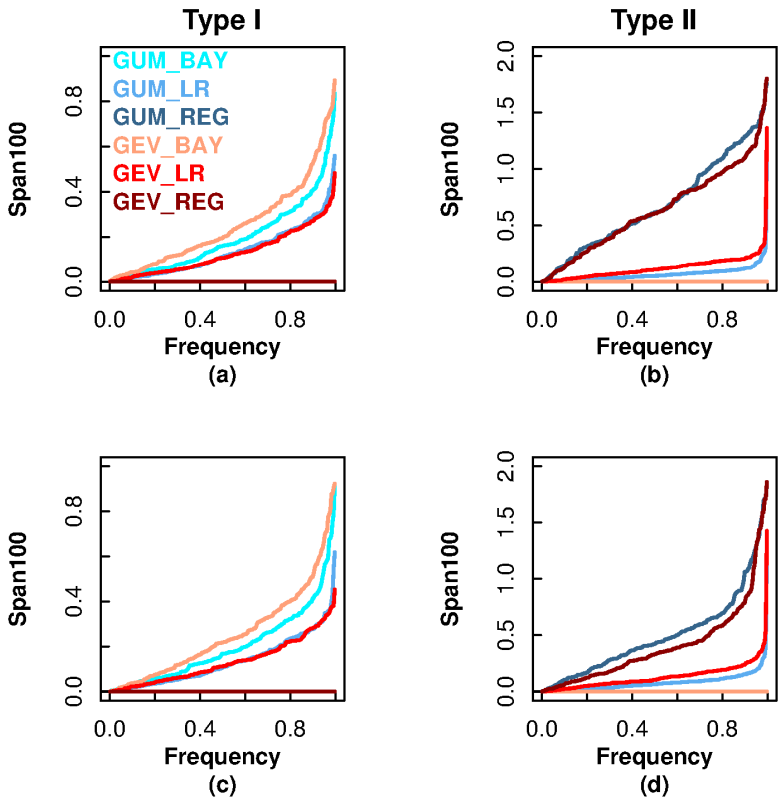


Figure 8. Stability diagnostic for six FA implementations (local, regional and mixed local-regional, with Gumbel and GEV distributions). Left = type I decomposition, right = type II decomposition. (a) – (b) = estimated distribution, (c) – (d) = predictive distribution.